Master Thesis

Identification of Business Travelers through Clustering Algorithms

By James Piggott

Capgemini

CONSULTING.TECHNOLOGY.OUTSOURCING

Master thesis Business and Information Technology

Faculty of Management and Governance
Faculty of Electrical Engineering, Mathematics and Computer Science

Identification of business travelers through clustering algorithms

Enschede, May 2015

Author
Name: James Piggott
Student number: s1011162
Email: jjhpiggott@gmail.com

Supervisory committee:
University of Twente
Chintan Amrit
Maurice van Keulen

External supervisor
Fai Greeve

Management Summary

Over the fiscal year 2014 Air France-KLM reported a net loss of €198 million (CAPA, 2015). Long-haul flights are a market in which the combined group has performed well in the past, yet it faces increasing competition (Skift, 2015). Efforts to compete on short-haul flights with established Low Cost Airlines (LCA) have proceeded slowly due to labour related problems with Transavia, the groups own LCA.

To better compete with LCA's, Middle East and Far East airlines as well as improve their operational profits Air France-KLM needs to better understand its passengers and their desires. Traditionally the business travel segment has been the group's most profitable segment. Previous market research has shown that only half such travelers have a corporate contract with Air France-KLM. This suggests that if Air France-KLM is able to identify which passenger is a business traveler it could improve its operational results.

Previous efforts to identify business travelers, performed by Arwed Wegscheid and Julia Godet at KLM, focused on using business rules, filters and data from questionnaires to create profiles. On the other hand, this method was not satisfactory because it relies on assumptions. Data clustering is about the discovery of any underlying structure within the data. The potential existed that flight movement data could not only discover new market segments but group passengers into those segments, which can be used for follow-up passenger targeting. The following research problem is answered in this thesis to discover if such a new method was feasible.

Design a new airline market segment model with data clustering

From a stakeholder analysis it is determined what the criteria are for the successful discovery of a new market segmentation model. The results of the data clustering were presented to a number of KLM customer management personnel who found them to be usable for future passenger targeting.

As a methodology to perform Data Mining research CRISP-DM was used. The sequential steps and guidelines of CRISP-DM acted as general guideline to perform this research effort in structured manner with the hopes that the process is reproducible to answer future business questions. The not inconsiderable effort to collect, collate and clean data from the Altea departure control system was worsened by the difficulty to find unique identifiers for passengers. Because unique identifiers are only kept for members of frequent flyer programs and corporate contracts it has been nearly impossible to tell passengers of the remaining group apart. A new unique identifier was constructed using first name, last name and Date of Birth (DoB). Despite gaps in the data because of a lack of DoB the approximation is faithful to the entire population of KLM.

Results Redacted

The results of this research are also applicable to other airlines if similar variable attributes are used. This offers the possibility for comparison of market segments and the potential for better market placement of airline offerings. Continued research should also focus on adding more descriptive variables regarding passenger tastes such as the sale of ancillaries and choice of in-flight options. However, the best way to leverage the potential of clustering algorithms is to ensure any dataset can uniquely identify a passenger.

Acknowledgement

This thesis is the capstone to my study Business & Information Technology. When mentioning this study and what it means in conversation I usually indicate that it bridges Computer Science, Industrial Engineering and Management. I have made every attempt to ensure that the research performed for this Master Thesis would reflect the cross-connectedness of these fields.

I must thank Fai Greeve, my external supervisor at Capgemini, and Matthijs Neppelenbroek from KLM for allowing me to investigate passenger segmentation with the powerful Machine Learning Algorithms. This master thesis uses a considerable number of techniques and algorithms and feels very much like a tour de force of this exciting field.

I would also like to thank Ajay Gopikrishnan and Willem de Paepe for their continued to support through the initial difficult time at Capgemini. Furthermore I would like to thank Arwed Wegscheid, Julia Godet, Maaike van der Horn, Sander Kempen and Lou Vermeer for aiding me. I should also want to thank the support I received from the many other people at Capgemini and KLM.

Furthermore I thank my supervisors at the University of Twente, Chintan Amrit and Maurice van Keulen, for their guidance. Especially Chintan, who has by all accounts spent a considerable amount of time with me free-thinking about the possibilities of this research.

This thesis marks the end of my college education, which I concluded at the University of Twente with more speed and fervor than I imagined possible. As a Science Fiction fan I am reminded that 'All Good Things End…' and I am now forced to look for other pastures, perhaps greener.

James Piggott

May 2015

Contents

List of figures

List of tables

Abstract

Over the fiscal year 2014 Air France-KLM reported a net loss of €198 million (CAPA, 2015). Long-haul flights are a market in which the combined group has performed well in the past, yet it faces increasing competition (Skift, 2015). Efforts to compete on short-haul flights with established Low Cost Airlines (LCA) have proceeded slowly due to labour related problems with Transavia, the groups own LCA.

To better compete with LCA's, Middle East and Far East airlines as well as improve their operational profits Air France-KLM needs to better understand its passengers and their desires. Traditionally the business travel segment has been the group's most profitable segment. Previous market research has shown that only half such travelers have a corporate contract with Air France-KLM. This suggests that if Air France-KLM is able to identify which passenger is a business traveler it could foreseeably improve its operational results.

If KLM can better understand the needs of their passengers they will be better able to target them with sharper pricing of tickets, increase client retention with frequent flyer programs and improve ancillary revenue. The goal of this research effort is to discover a better market segmentation model suitable for use by Air France-KLM.

In order to achieve the goals of better understanding airline passengers this research effort uses actual passenger flight movement data. Records were collated and grouped such that for each identifiable unique passenger a record exists of all their flight movements within a year. Metrics such as frequency of travel, distance traveled and the weight of baggage checked in are 3 among more than 25 variables that permit passenger behavior to be identified and passengers grouped together.

This research effort used machine learning techniques aimed at recognizing airline passenger behavior and grouping them together. The theory behind techniques such as supervised and unsupervised learning is discussed. Practical issues with unsupervised learning algorithms such as K-means, Expectation-Maximization and Hierarchical Clustering Algorithms are also detailed. Such algorithms make it possible to group data points that share similar features together. Such clusters are then identified according to existing airline market segments and interviews with stakeholders.

For supervised learning algorithms such as Decision Tree (CART) is explained. Supervised learning makes it possible to create a predictive model of labeled data. It is thus possible to describe when an airline passenger, as an example, should be recognized as a business traveler.

This thesis expands this effort by also using semi-supervised learning. This technique has the advantage in that it requires only a sample of the data to be labeled for use in training an algorithm. Semi-supervised learning offers the possibility of being more accurate than unsupervised learning without having to incur the cost associated with labeling the data set.

1. Introduction

1.1. Problem statement and Background

To better compete with Low Cost Airlines (LCA) and to improve their operational profits Air France-KLM needs to better understand its passengers and their desires. If this can be achieved they will be better able to target passengers with sharper pricing of tickets, increased loyalty to frequent flyer programs and improve ancillary revenue. The goal of my research is to continue efforts of better defining the passenger segments of Air France-KLM. The need to identify business travelers and the possibility of using clustering algorithms on travel means the result could provide a unique solution. The results may help Air France-KLM with better understanding their passengers and accommodate their needs.

1.1.1. Goal

This thesis summarizes efforts to use KLM data on passenger movements to conduct clustering analysis. A specific focus was placed on identifying uncontracted business travelers in the hopes to learn more about their demographics and behavior. Business travelers in general form the most profitable passenger group. Past research has proven the possibility of using clustering techniques to identify market segment, especially high-value customers with the intention of adapting retention strategies (Maalouf, 2007). A previous cursory study performed internally at KLM suggests that many passengers show similar behavior and traits to business travelers but they have not been identified as such and are not enrolled in Flying Blue or BlueBiz, the KLM corporate frequent flyer program for small and medium sized businesses, or other corporate programs. This group of KLM travelers has not accepted special business offers, thus reducing KLM's potential revenue from this client pool. Previously efforts were made to better determine the size of this group by applying expert knowledge, business rules and filters obtained from questionnaires. Factors such as age, gender, distance traveled in a year and corporate email address were applied to segment travelers into groups. When known contracted business travelers were subtracted a large pool of potential clients remained.

1.1.2. Scope

Through this thesis readers learn about efforts to apply machine learning algorithms to cluster passengers according to data obtained from KLM bookings and passenger movements. From the results customer segments and behavior are inferred. These results are validated by comparison to internal KLM market analysis and through interviews with experts. The cluster analysis would be conducted as an unsupervised machine learning problem using algorithms such as K-means, Hierarchical Clustering and density based algorithms. The travel industry already has a long history of having used clustering to identify market segments, this research would attempt to provide readers with a thorough overview of the algorithms used and determine how the connection is made between an established segment and knowledge gained from clustering. After segments have been identified and validated they are used as prior knowledge as part of supervised learning. The results of give an indicator of the relationship between business travelers and leisure travelers. The intention is that the methods of this study are generalizable to other businesses.

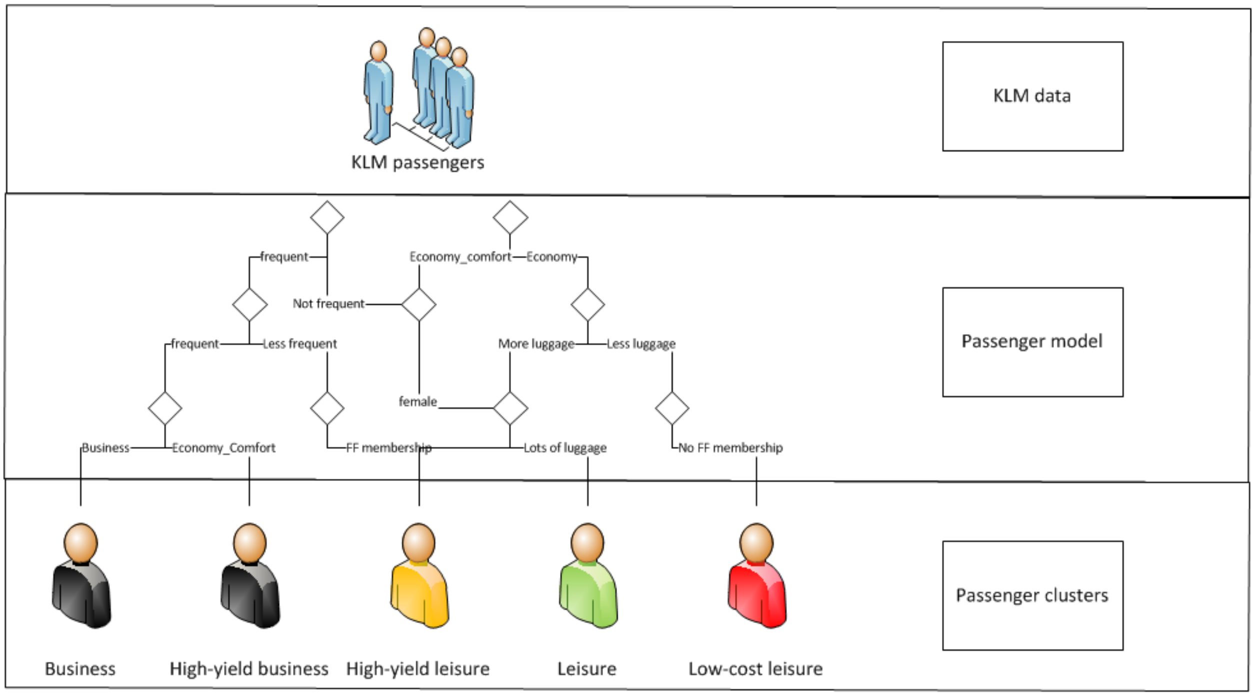


Figure 1 The potential of data clustering with KLM passenger movements

1.2. Motivation for using Machine Learning

1.2.1. Clustering airline passengers

Cluster analysis is a subfield of Machine Learning that also includes but is not exclusive to: Neural networks, Decision Tree Learning and Association Rule Learning. Cluster analysis is usually defined as an unsupervised learning method, in that no labelled data is available to guide the learning method to its result. The problem this research faces is that airline passengers that show similarity with business travelers cannot be identified as such because they are not labeled and are not known as business travelers. There exists precedent for the use of cluster analysis in the tourist industry (Jain 2010 and Brida at. Al, 2014), customers are grouped according to preferences in order to better accommodate them.

1.2.2. Theoretical basis of data analysis

According to Tukey (1977) data analysis can be broadly grouped into exploratory and confirmatory categories. The former is descriptive in nature which means that the investigator does not have pre-specified models or hypotheses but they want to understand the general characteristics or structure of high-dimensional data. Confirmatory or inferential deals with an investigator seeking to confirm the validity of hypothesis, models or a set of assumptions. Before the research is completed it is impossible to say whether data analysis and clustering will succeed and produce results of interest for a business. The usefulness of any data-driven segment identification depends on two things: the quality of the data and the best possible use of the explorative tool of cluster analysis (Dolnicar, 2002).

1.2.3. Previous research

Previous research by Jain (2009) defines the purpose clustering as: “the use of clustering to find structure in data is inherently explanatory in nature. It is a formal study of methods and algorithms for grouping, or clustering objects according to measured or perceived intrinsic characteristics or similarity”. This research builds on previous efforts to gain insight into passenger behavior. Previous efforts at KLM have used data obtained from questionnaires to build personas that reflect the greater population from which airlines obtain its passengers, but such a model has proven unwieldy and too reflective. Another effort focused specifically on the characteristics of uncontracted business travelers derived from a set of assumptions made by an internal market analysis, this effort to have a narrow focus to prevent overlap of passenger behavior and relied too much on expert knowledge which may be biased. However, this second effort was the catalyst for this research and it hopes to emulate its finding through clustering. Both previous results will be used as a benchmark.

1.3. Research questions

The main research question of this master thesis is:

Design a new airline market segment model with data clustering.

For the purpose of data clustering all passengers movements recorded over 2014 in Altea Departure Control (Altea_Dc) are used. These records are grouped according to each uniquely identifiable passenger in the hopes that their behavior patterns will allow them to be group by clustering algorithms. This research question is divided into three sub questions that combined address the goal of this research.

Can clusters be associated with passenger segments and types?

This sub question is answered by performing various unsupervised learning algorithms on the airline data set. The resulting clusters will hopefully resemble different types of passenger segments. To prevent the clusters from being artifacts of the algorithm various popular algorithms are used and their results compared. These algorithms include K-means, Expectation-Maximization and Hierarchical Clustering. The results are evaluated by submitting them to airline stakeholders to discover whether passenger segments can be identified. Validation is performed through analysis of various metrics such as Sum of Square Error (SSE), the results from the clustering are also validated by treating them as prior knowledge of classification problem.

Can an airline’s existing practice of customer segmentation be improved?

This sub question describes a situation where an airline, or an any business or organization, already has a customer segmentation and has classified a number of passengers to correspond to a segment. As new customer records are added continuously over the life of a business it should be possible to classify them without performing unsupervised clustering as was done with the first sub question. With semi-supervised learning the records that already have been labeled will be used to guide the learning process and thus help classify unlabeled records automatically. The ability of having no longer the need to manually classify all passenger records and the ability to shift segments with changing behavior makes this application of semi-supervised learning desirable for large businesses.

Can behavior of airline passengers be modeled?

To answer this sub question supervised learning is used to create a model of each cluster identified in the first sub question. Such models are created using Decision Tree learning. The cluster that has been identified is a variable in the data set hence the term supervised learning because the result are already known. The term classification or regression learning also often used. The results are expanded by attempting to explain why passengers have a corporate contract or are member of the KLM frequent flyer program.

The combined results of all three sub questions are presented to the principal stakeholders to determine their possible utility and thereby their validity. A cursory observation of the literature suggest that there have been similar scientific studies, which this proposal could expand (Mahrsi et al, 2014 & Brida et al, 2014). Segmentation of customers is an important business strategy, but it has been superseded by attempts to fulfill customer needs on an individual basis through micro-segmentation and personalization (Huls et al, 2014), but has not been universally successful.

1.4. Methodology

According to Wieringa (2014, P.3) "Design science is the design and investigation of artifacts in context". The artifact, which can be algorithms and methods do not solve any problem, but their interaction with the problem context does. Before this research effort can proceed with applying the many Machine Learning algorithms and Data Mining methods that are available, the goals and success criteria should be determined so they can act as a guideline.

To ensure that the right procedures are applied the correct research type needs to be identified. After this is accomplished the successive steps to be taken to answer the research questions can be followed. This chapter concludes by presenting a research model.

1.4.1. Research types

Wieringa (2010) identifies two types of research problems, design problems and knowledge questions. Each require their own kind of research questions to delineate and problem-solving cycle to answer. Both types of research problems are described as follows:

- **Design problems** call for a change in the real world and require an analysis of actual or hypothetical stakeholder goals. The solution is a design, and there are usually many possible solutions. Design problems are evaluated by their utility with respect to the stakeholder goals.
- **Knowledge questions** ask for knowledge about the world as it is. The assumption is that there exists only one correct answer. The answer may be incorrect, or have a degree of uncertainty but the answer is evaluated by truth, it does not depend on stakeholder goals.

The distinction between the two is often camouflaged in research by the way research questions are formulated. The main research question discussed in this thesis is '*Design a new airline market segment model with data clustering'* can also be rephrased as '*Can a new airline market segment model be created with data clustering'*. To answer the rephrased question would be a lot easier. The answer can be yes or no or either with a degree of uncertainty. However, this question and its answer would not attain the stakeholder goals. To attempt to answer the main research question a stakeholder analysis is performed and the results of this research are evaluated through interviews with the stakeholders.

As problems can create new problems and questions a design science project is never restricted to one kind of research problem. A research project can iterate numerous times through the problem-solving cycle. This research effort starts with a design problem which in turn raises three knowledge questions. The first sub question '*Can clusters be associated with passenger segments and types?'* is evaluated by truth. The answer leads to the second sub question '*Can an airline's existing practice of customer segmentation be improved?'* which in turn leads to the third sub question '*Can behavior of airline passengers be modeled?*' before the results of all three are evaluated with the stakeholders to answer the main research question.

1.4.2. Research steps

Now that the research questions have been identified as one design problem and three knowledge questions the proper successive steps to answer all three can be planned. Both types of questions have different problem-solving cycles although they do share similarities. The main research question is answered using the engineering cycle. It describes a set of tasks that are logically structured in an attempt to make an improvement for stakeholders in a rational way. The four stages of the engineering cycle are as follows (Fernández and Wieringa, 2013).

1. **Problem investigation:** the stakeholders are identified as well as the goals they have. Furthermore practical phenomena that exist are investigated and the effects they have and what it means for the project goal contribution.
2. **Treatment design:** the first design choice is made, in this case the specifications of requirements for a treatment. Also the available treatments are investigated.
3. **Design validation:** the design must be investigated to see what its effects are and whether it will satisfy the requirements. Alternative treatments must be considered (trade-off analysis) and sensitivity to changes in the problem context must be investigated.
4. **Treatment implementation:** this entails the transfer to practice after which the treatment has been realized and is outside the control of the designer of the treatment.
5. **Implementation evaluation:** determine how successful the treatment has been.

In this context the word treatment would otherwise mean the same as solution, but as the artifact may only partly solve the problem the word treatment is more suitable. Each of the five steps of the engineering cycle is answered by knowledge questions. This research effort attempts to traverse the engineering cycle by answering three sub questions which have their own problem-solving cycle, the empirical cycle.

1. **Problem investigation:** determine what the research problem is that needs to be solved.
2. **Research design:** determine what needs to be done to solve the problem. It concludes with inference design to determine how to draw conclusions generated from data.
3. **Design validation:** match research design with inferences from the data.
4. **Research execution:** research is executed. Events relevant for the interpretation of results must be reported.
5. **Results evaluation:** determine if there is anything that remains to be solved.

With the empirical cycle a considerable level of latitude in deciding to what detail each step is performed. The cycle also does not need to be followed in the sequence described above. Below the results are summarized in a graph to convey how design science is relevant for this research effort. The overall guiding cycle is the engineering to answer the principal research question. The empirical cycle is performed three time to answer the sub research question. Their results influence the second, third and fourth steps of the engineering cycle while their answer, validated based on truth, would be inconsequential to the fifth and final step, implementation evaluation. The next paragraph, on thesis structure, explains how this series of cycles recurs in the remainder of the thesis.

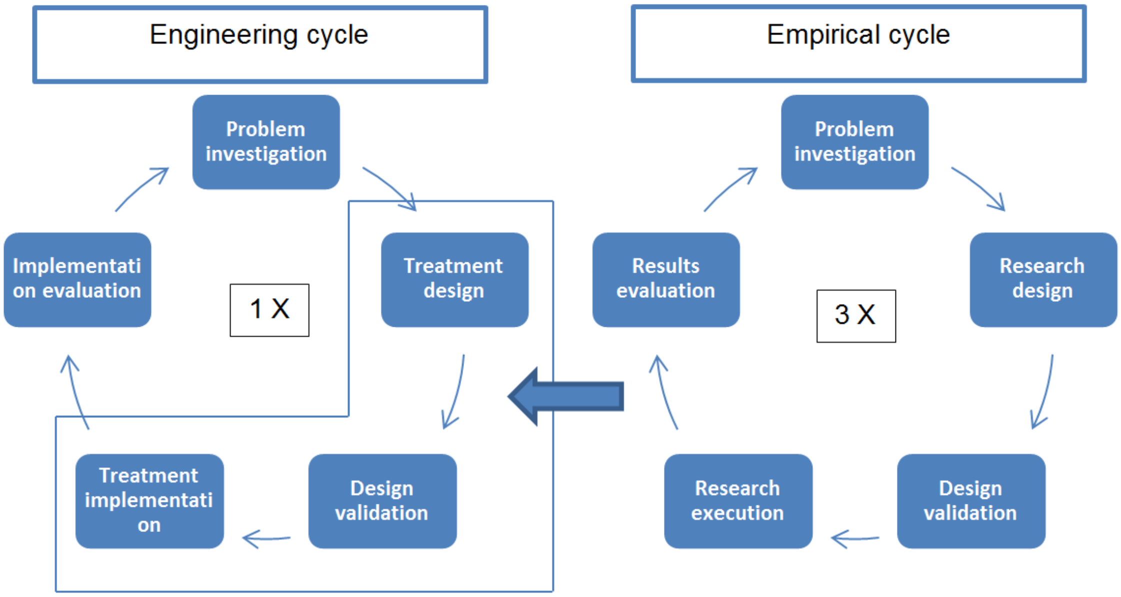


Figure 2 Proposed problem-solving cycle

1.5. Thesis structure

This section gives an overview of the structure of this thesis and how it corresponds to the proposed problem-solving cycle.

Thesis chapters:

- Chapter 1. Introduction
- Chapter 2. Background and stakeholder analysis
- Chapter 3. Structured Literature Review
- Chapter 4. Machine Learning in Business Analytics
- Chapter 5. CRISP-DM methodology
- Chapter 6. The KLM Dataset
- Chapter 7. Results
- Chapter 8. Evaluation with stakeholders
- Chapter 9. Conclusion

The first step performed in any research is a small literature study and stakeholder analysis to confirm the research objective is viable. Stakeholders were asked whether there was access to data and whether the data set can in theory provide information to distinguish airline passengers. The small literature study was to discover whether there was a precedent and determine if this research effort could possibly add to established knowledge. This step was capped by a research proposal. The result of this conforms to the 'problem investigation' step of engineering cycle and form most of the material found in chapters 1 and 2.

The second step was in-depth study of established literature using the Structured Literature Review approach proposed by Kitchenham (2004). The results of this are found in chapter 3 and correspond to the first step of the empirical cycle.

The third step consisted working with the data set and transforming it to a format that enabled analysis. As more variables were finalized the data set was tested using small scale clustering tests. This step was only finished when the list of possible variables that could be useful was exhausted. This step corresponds to the remainder of the empirical cycle: research design, design validation, research execution and result evaluation and can be read in chapter 4 through 7.

The fourth step consisted of a stakeholder analysis that was more in-depth than the questioning that was performed surrounding the initial research proposal. In this step the small tests from step 3 are used to clarify possibilities to every independent actor involved in the research effort. This step finalizes what the main deliverables of the research are and whether its results attained stakeholder goals. It corresponds to the final phase of the engineering cycle 'implementation evaluation' and can be read in chapters 8 and 9.

2. KLM Background and Stakeholder Analysis.

2.1. KLM background

KLM is the world's oldest airline established in 1919 while Air France was established in 1933. The Air France–KLM group was established through a merger in 2004, after which KLM became a member of SkyTeam. Despite the merger KLM retains much independence, it has its own headquarters in Amstelveen, Netherlands separate from the group's headquarters in Montreuil, Paris. Combined the Air France–KLM group transported 78.45 million passengers in 2013.

Despite KLM's independence many projects seek to establish better synergy between the airlines to further improve its financial health. One such joint project is the replacement of the departure control systems Gaetan (Air France), Corda (KLM) and its check-in module Codeco. The Corda departure control system dates from 1969. The replacement for these systems, Altea Dc has been co-developed by the Spanish Amadeus IT Group. The system, once completed, offers end-to-end customer experience that is fully automated (Amadeus, 2015). The Amadeus reservation system was implemented at KLM starting in 2007, despite the fact that Altea has been in use at over 200 other carriers the version in use by Air France-KLM will be fully customized (Altea DC News, 2011).

2.2. KLM Frequent Flyer programs and contracted customers

Air France-KLM maintain several loyalty programs for those who travel frequently for business or other purposes. Business travelers may receive discounts if the company they work for has established contracts with KLM. Below is an overview of such loyalty programs.

2.2.1. Flying Blue

The Frequent Flyer program for Air France-KLM is Flying Blue. It allows members to collect Miles which can later be spend on upgrades, free flights or gifts. The program makes a distinction between Award Miles and Level Miles. The former works similar to other frequent flyer programs and client retention schemes as it allows users to purchase with them tickets, products and services at well over a 100 partners. Level Miles allow users to obtain a higher Flying Blue participation level and receive better services and promotions. It also increases the rate at which Award Miles are earned. Flying Blue has all the hallmarks of a classic client loyalty scheme that now also has features similar to that of a credit card service. Flying Blue has four participation (tier) levels: Ivory, Silver, Gold and Platinum. The latter three categories are considered part of the Elite status that passengers can obtain. For clients to be awarded Miles they only need to either show their membership card or fill in their membership number. In general, every mile a passenger fly earns them one Flying Blue Mile. Miles can also be accrued by flying with SkyTeam partners.

For passengers to obtain a higher tier level they either need to fly more than a set number of Qualifying Flights or earn more than a set number of Level Miles. Only Ivory tier has no such requirement. If passengers have performed more than the required number of Qualifying flights or Level Miles but have not reached the requirements for the next higher tier than the difference will carried over into the next year. On December 31st of every year the Annual Level Check is performed. For passengers to keep their membership level they will to reach the required threshold of Qualifying flights or Level Miles every year.

Membership level	Required number of Miles	Required number of Qualifying flights
Ivory	Entry level	Entry level
Silver	25,000 Level Miles	15 qualifying flights
Gold	40,000 Level Miles	30 qualifying flights
Platinum	70,000 Level Miles	60 qualifying flights

Table 1 Frequent flyer membership conditions

Other programs

Flying Blue also support various special level groups targeted at specific demographics.

1. **Jeune** is a program aimed at youths aged between 2 and 24 who are residents of Metropolitan France, one of its overseas departments or are resident in Morocco, Algeria or Tunisia.
2. **Petroleum** is aimed at those working in the Petroleum industry. Membership offers similar perks as other Flying Blue tier levels but are only available on predesignated oil routes. Members must already be enrolled in Flying Blue to participate while Jeune members are excluded.
3. **Seamen program** is aimed at those working at sea who need to frequently travel to other port cities.

2.2.2. BlueBiz

BlueBiz is another frequent flyer program specifically catering to business travelers working for small and medium sized firms (< 150k revenue). BlueBiz Credits can be accrued simultaneously with Flying Blue which can then also be redeemed for free flights, upgrades and a wide range of items. BlueBiz Credits are for use by corporations and only they can bestow perks to employees. BlueBiz is thus distinct from Flying Blue in that it is a company loyalty retention program.

2.2.3. Business Contracts

There are also business travelers whose companies have contracts for discounts with KLM.
The deal is dependent on the number of flights made by such company employees, the more flights the higher the discount that can be arranged. Important accounts have a revenue in excess of 1 million per annum. For regional flights this is between 150 K and 1 million.

Because of the information gathered through the above mentioned loyalty programs KLM knows more about such passengers than they do about travelers that do not take part.

2.2 KLM, market segmentation and behavior

Market segmentation is a strategy of dividing a broad target market into subsets of clients who have common a characteristics and needs (Haley, 1968). The purpose of market segmentation is to be able to adapt your marketing efforts as well as your product offerings to better suit potential customers and thus raise revenue. Companies are challenged by finding better suited segments. There is a slow drift from market segmentation towards (micro)-segmentation. Companies are also driven to provide tailored interaction with their customers to provide optimal service, this is referred to as personalization. The potential of market segmentation is obvious: targeting a market segment characterized by expectations or preferences leads to competitive advantage. With cluster analysis it should be possible through an explorative analysis to identify the KLM's market segments.

KLM has already conducted numerous market research projects. The results of two can be considered the catalyst of this research. The following two sections give an overview of each.

2.2.1 Market personas

The first research results pertinent to this investigation was the creation of market personas based on extensive questionnaires. The results were summarized in 7 persona's: 3 business and 4 leisure.

Table Redacted

The market personas represent the larger population that KLM draws its passengers from. Despite enthusiasm for these personas they are not universally applied. Recent interest on further micro-segmentation (personalization) of passenger contact was cited as one factor.

2.2.2 Generalized model of KLM business traveler

The model below of business travelers has been developed at KLM by Arwed Wegscheid and Julia Godet using expert knowledge, business rules and questionnaire responses. This model signifies what is possible and gives an indication of the direction this research should take. This information can be used to compare results found in data clusters to proof the 'external validity' of the research questions. Not all of these criteria are mutually exclusive though, any set of combination should be regarded.

Business model Redacted

2.3 Stakeholder analysis

The goal of this stakeholder analysis is to identify all biological or legal persons affected by the proposed artifact that stems from the design problem. Using the stakeholder onion model developed Alexander (2004) an attempt is made to determine how close stakeholders stand with regards to the artifact and what their role is in its context. In doing so their relationships becomes clear. For each stakeholder a brief description is given in section 2.

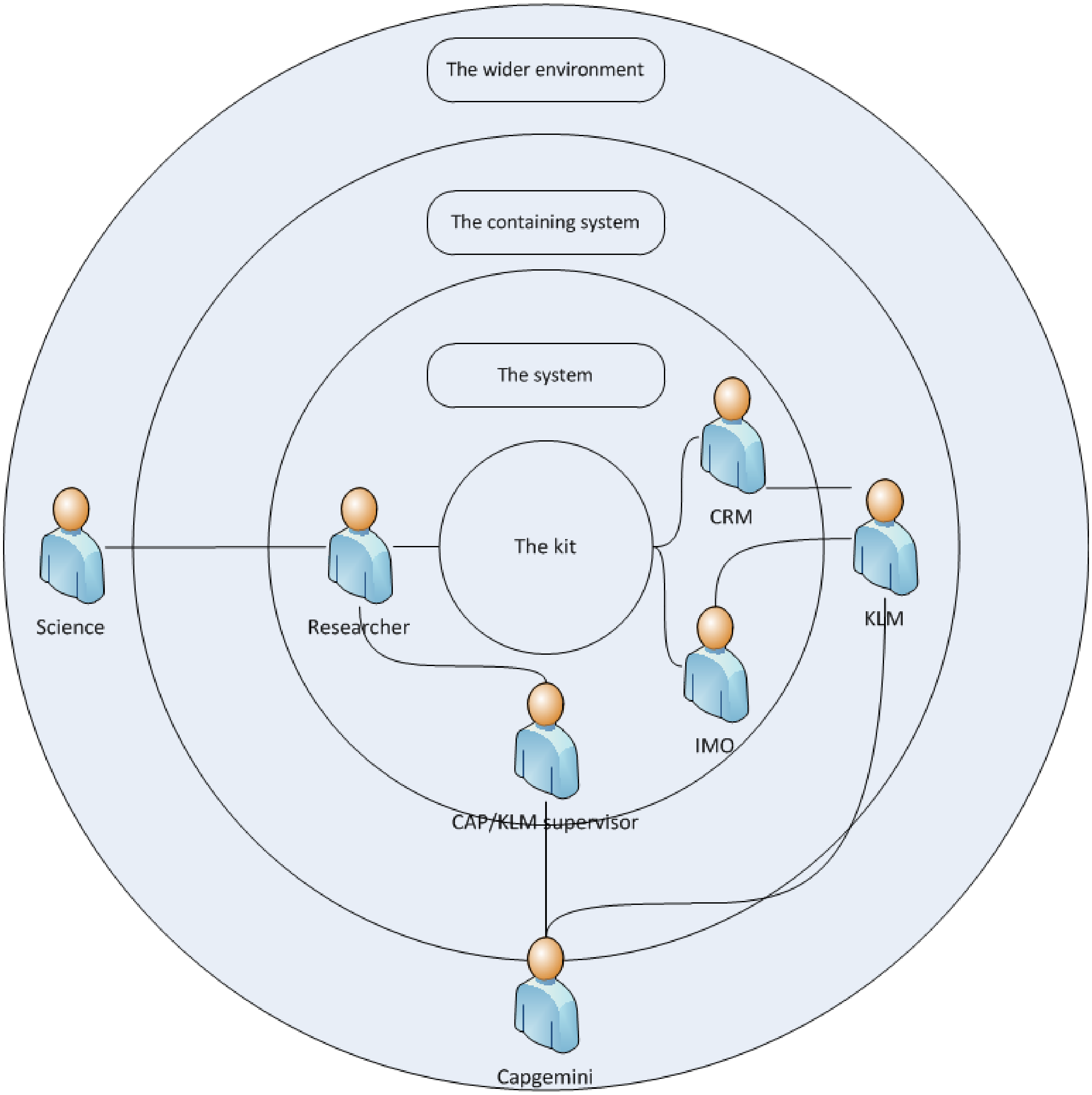


Figure 3 Onion model of stakeholders (Alexander, 2004)

In the onion model by Alexander and Robertson the level of interaction that stakeholders have with the artifact (the kit) is depicted by whether not they have direct access, or come into contact with the artifact through other systems or are part of the larger environment in which the artifact is placed. Each stakeholder also performs a particular role. In this case stakeholder roles range from developer (researcher), clients (CRM department), sponsor (CRM, IMO and Capgemini/KLM supervisor), functional beneficiaries (KLM), negative stakeholder (staff with superfluous skills), political beneficiary (IMO) and financial beneficiary (Capgemini and KLM). Some stakeholders can have two or more roles, but for clarity and brevity this has been reduced as much as possible. The relationships between stakeholders is also depicted by the lines between each figure in the graph. Again for clarity and brevity only the principal relationships are depicted.

2.3.1 Stakeholder description

KLM

Through improved customer satisfaction KLM aims to strengthen its competitive position in the airline market. A better passenger model can ensure more competitive pricing. This will increase KLM's position vis-à-vis close partners and competitors. A passenger model that can be updated automatically to adjust to changing passenger behavior and market circumstances may allow for further cost reduction through employee downsizing.

Customer Relationship Management

The task of Customer relationship management (CRM) is to manage interactions with current and future customers. It applies to wide variety of topics including sales, marketing, customer service, and technical support. CRM desires to leverage customer information flows from all contact points to provide better service. A detailed model of passenger behavior can be used to adjust their services to suit individual passenger needs, thus improve how they already interact with passengers. CRM desires to move away from market segmentation towards customer personalization. Customer support already uses social media to answer passenger questions and inform them of schedule changes. However, as previous market analysis used information obtained from questionnaires and online contact points this study has the potential to create a fundamentally new passenger model based the behavior of all passengers. Such a model can be used to anticipate passenger whishes as well as act as a benchmark for customer interaction.

Information Management Office

The task of the Information Management Office (IMO) is to align people, process, and systems. Too often value associated with innovation or deals fails to materialize because the integration strategy is not properly applied. IMO's goals of this research is to develop the outcome and ensure it potential is fully realized.

Researcher

The first goal of the researcher is to complete the research questions in an academically acceptable way that will ensure graduation. The internship that is associated with the research has the goal of providing practical experience to improve practical skills and future employment. The researcher has a lot to gain by making sure the design artifact is used in practice.

Science

The field of Data Analytics would stand to gain to discover whether theoretical methods of data clustering can be proven to work in practice.

Capgemini

The successful outcome of the project for KLM has the potential of increasing business for Capgemini. The goal of improving passenger modelling can be considered a case study that may be applicable to other situations. A successful outcome will increase the knowledge base for Capgemini that may be leveraged to increase business.

Supervisor

The supervisor is an employee of Capgemini working as external consultant at KLM. The research goals are important to ensure future employment for Capgemini. Successfully guiding the researcher towards attaining the research goals will ensure a gain in perceived management skills.

2.3.2. Stakeholder Awareness

The final part of the stakeholder analysis is to determine how aware stakeholders are of the research being carried out and what resource capability they have to offer (Wieringa, 2012). Each research effort needs to be assigned a budget in terms of resources. Stakeholder resources need to be assigned to find a solution to the problem. Whether they can allocate resources depends on their awareness. There are three possibilities.

1. **Not aware;** stakeholder is not aware of a treatment nor sees the need for one. An event pushes the possibility into awareness.
2. **Passively aware;** The stakeholder is aware of the possibility of a treatment but does not consider it important enough to do something. An event ensures that the stakeholder makes resources (time, money) available.
3. **Aware & committed;** resources are committed to act to attain a goal.

The prior research carried out by business analyst Arwed Wegscheid and Julia Godet raised the possibility of creating a market segmentation model to identify business travelers through data clustering. However, the principal stakeholder has been Matthijs Neppelenbroek who is a project manager for IMO. His suggestion for a research project ensured the availability of resources. With the start of the research project, after the proposal was approved, other stakeholders became aware of the possibilities of a treatment through interviews. This ensured time with expert knowledge and access to data were made available. Their role changed from being only passively aware to becoming aware and committed to the treatment.

3. Structured Literature Review

3.1. Literature Methodology

In this chapter the methodology behind finding and processing the literature is explained. The scope, inclusion and exclusion criteria and the results are discussed. Following such a pattern is called Systematic Literature Review (SLR), guidelines for which were obtained from Kitchenham (2004). The search for literature focused on two sources: Scopus and Google Scholar. Scopus offers the possibility to narrowly define the search criteria for literature while Google was used to broaden scope and find papers that may have been missed by Scopus. A common problem is the use of synonyms in the field of data clustering. By scanning the abstracts of papers that met the criteria those that were not deemed relevant were excluded. The principal criteria for selections is whether a paper will probably help in understanding the problem and answer the research questions.

3.2. Scope of SLR

To narrowly define the search criteria to complete this literature review is almost impossible without a process of 'trial and error'. Overarching terminology such as Data Mining, Machine Learning and Clustering mean different things to different people working in these fields. Data Mining is too broad as it includes Artificial Intelligence, Machine Learning and Statistics. Machine Learning is a broad terminology used for Regression Analysis, Clustering, Neural Networks and Classification (supervised learning). To perform a SLR according to the principals and guidelines laid out by Kitchenham (2004) a broad set of papers regarding Data Mining, Machine Learning and Clustering is explored to better define the scope and search terms. Dolnicar (2002) gives an overview of previous efforts to define new market segments based on tourist information through data clustering. It compares studies based on the method of algorithm, the criteria for validity and selection of appropriate variables. Two other defining articles are by Jain (1999, 2010) which describe both a wide range of applications for clustering and suitable algorithms.

3.2.1. Inclusion criteria

This study includes journal papers and in some cases also conference/workshop papers when there too few results. The scope did not extend to editorials, letter to authors, summaries of discussions and so on. All texts are written in English. Terms such as clustering, classification, cluster validity, CRISP-DM, Machine Learning and Semi-supervised learning were included in the search criteria. The latter term was unknown to the researcher. It was discovered that because of the broad meaning of some terms there can be misunderstandings. Initially the terms classification and supervised clustering (not supervised learnings) were used interchangeably. However, the latter term only applies to validation of clustering results by treating it as a classification problem.

3.2.2. Exclusion criteria

This study excludes papers that try to proof the correctness of algorithms with math's, or evaluate variants of algorithms that are tailored to a specific problem domain. Initially the search for literature sources focused solely on Scopus due to its options to limit results according to certain criteria. During the latter the stages of this research the literature sources were expanded to include Google Scholar for those topics that did not yield fruitful results or regarded minor issues. The search criteria and boundaries for Scopus were: (1) limit publications to journal papers instead of conference papers if possible (2) limit papers to those written in English (3) limit publication date between 2004 and 2014 (4) the subject area must include computer science in order to be relevant for the scope of this research.

3.2.3. Search terms and Query

An example of a search query using Scopus and the term clustering looks as follows.

TITLE-ABS-KEY (**clustering**) AND PUBYEAR > **2003** AND (LIMIT-TO (DOCTYPE , **"ar"**)) AND (LIMIT-TO (SUBJAREA , **"COMP"**)) AND (LIMIT-TO (LANGUAGE , **"English"**)) AND (LIMIT-TO (SRCTYPE , **"j"**))

This yielded 18.131 journal papers. Far too many to filter by reading abstracts. In this case a more pragmatic approach was taken by selecting those with a high number of citations. For this paper 16 articles were used. Other terms such as Data Mining and Machine Learning were treated similarly. The Data Mining method CRISP-DM and clustering topics such as semi-supervised learning had manageable results. The term CRISP-DM returned 31 results. 6 of these papers discussed knowledge discovery in fields such as medicine, after a cursory reading of their abstracts were discarded. Overall the volume of work was disappointing. After reading the abstracts 25 were deemed relevant of which 8 were only partly relevant. Of 11 articles no free copy could be found. Of the remaining 14 articles 5 were used. A similar procedure yielded 7 papers on semi-supervised learning. 8 more papers were applied after filtering results of a search using Machine Learning as a search term.

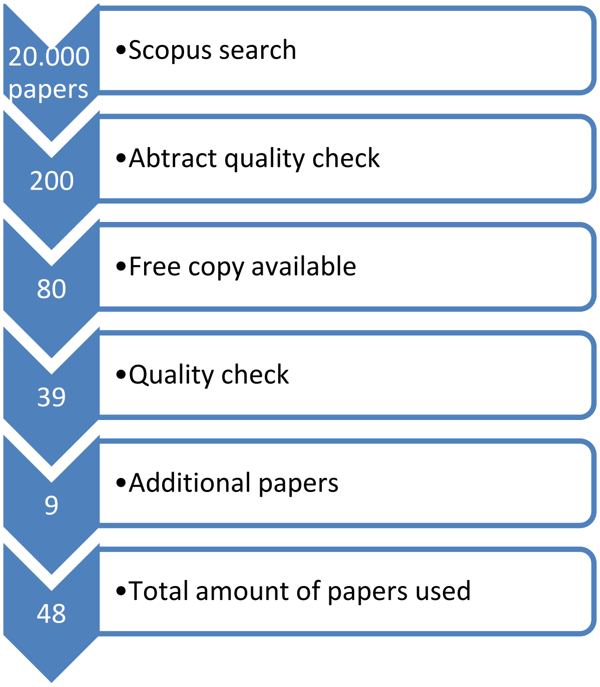


Figure 4 Results of Structured Literature Review

3.3. Additional articles

A number of research papers were added by using the Google search engine without use of inclusion or exclusion criteria or were added after recommendation by fellow researchers. Topics that were searched included ‘design research’, ‘micro-segmentation’, ‘stakeholder analysis’ and ‘cluster validation’. Together 9 more papers were added.

3.4. Books sources

Additionally to scientific papers three books describing topics on Machine Learning and clustering were used. *Machine Learning: An algorithmic Perspective* by Marsland (2nd edition 2014) was used to obtain a greater knowledge of the workings of various supervised and unsupervised learning algorithms. *Machine Learning: Hands-On for Developers and Technical Professionals* by Bell (2014) proved to be a practical source for the implementation of algorithms through libraries found in R and in WEKA, this book filled a knowledge gap that no scientific paper had an answer to. Additionally chapters on clustering and classification from *Introduction to Data Mining* by Tan, Steinbach and Kumar (2006) were used to gain an overview of those topics. A fourth book, by Roel Wieringa, entitled ‘*Design science methodology for information systems and software engineering*’ was used to organize the steps taken to complete this research in an logical and scientifically justifiable method.

4. Machine Learning in Business Analytics

This research effort relies heavily on Machine Learning algorithms, especially those for clustering, to describe passenger behavior and create a model that can be generalized. To exclude the possibility that clustering results may be an artifact of the algorithm used the focus will be on the application of different types of algorithms. Only then can the best possible use of explorative tool of cluster analysis be determined. In the next section the difference between supervised and unsupervised clustering is described and why each has different algorithm. The section afterwards describes in more detail unsupervised clustering algorithms.

4.1. Supervised and unsupervised clustering

According to the definition of Grira et al. (2005) “Clustering (or cluster analysis) aims to organize a collection of data items into clusters, such that items within a cluster are more “similar” to each other than they are to items in the other clusters. This notion of similarity can be expressed in very different ways, according to the purpose of the study, to domain-specific assumptions and to prior knowledge of the problem“.

Clustering algorithms can be divided into two groups: supervised and unsupervised. The former is a recent addition and makes use of small amounts data that are already classified (the supposed end result is already known) to infer a model or function. Examples of commonly used algorithms are Decision Tree learning, Artificial Neural Networks and Bayesian Algorithms. Unsupervised algorithms are used when no information is available concerning the membership of data items to predefined classes (Grira et al, 2005).

4.2. Unsupervised Clustering algorithms

If you were to boil down all the definitions of clustering, you get ‘organizing a group of objects that share similar characteristics’ (Bell, 2014). Such groups or clusters need to be part of the underlying structure of the data and not artifact of the algorithm. Finding clusters in data is easy for humans, if the dataset can be visualized in a 2 or 3 dimensional plane. For higher-dimensional data clustering algorithms are needed. The goal of such algorithms is to identify a number of clusters that would represent the structure of the data. In the case of partitioning algorithms such as K-means the number of clusters needs to be known a priori. There are many unsupervised clustering methods. Due to the strong diversity of the existing methods, it is impossible to obtain a categorization that is both meaningful and complete. Jain et at. (1999) come close in their seminal work on data clustering. Grira et al. (2005) expanded their attempt at an taxonomy of methods delivering the following results.

- Partitional clustering
    - Methods using the squared error
    - Density-based methods
    - Mixture-resolving
- Hierarchical clustering (dendrogram)

Chaudhari and Parikh (2012) consider density-based methods as a separate class due to the algorithms need for density drops to detect clusters. Each of the categories is discussed below; the comparative advantages of use and speed are also explained.

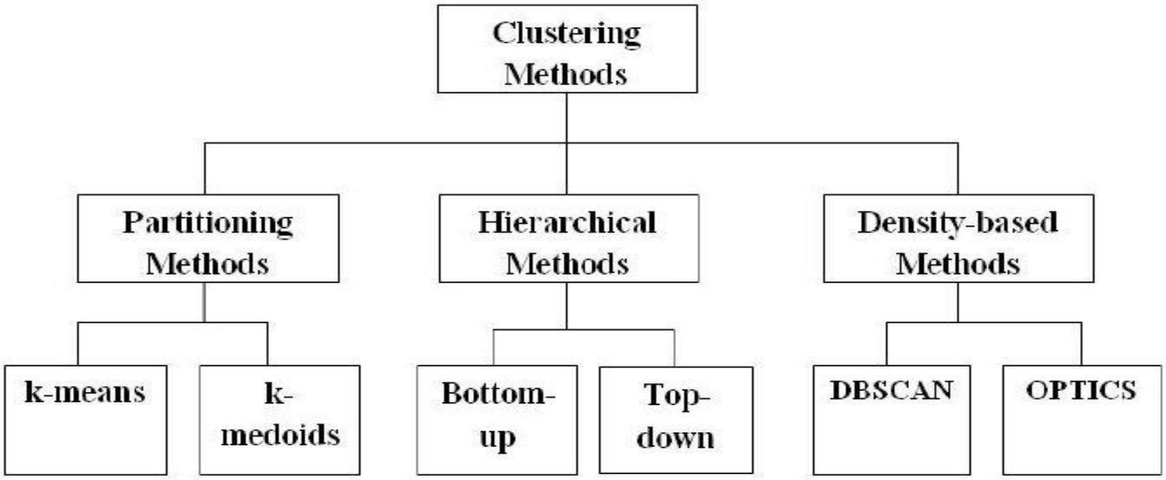


Figure 5Taxonomy of Chaudhari and Parikh

Partitioning clustering

K-means clustering

K-means is probably the oldest clustering algorithm still in use. Originally devised in 1957 and implemented in 1967 by James McQueen. It also known as Lloyd's algorithm in the field of Computer Science (Lloyd, 1982). K-means clustering is a partitioning algorithm with the objective of grouping a set of dissimilar data points into disjoint clusters. K-means attempts to minimize the within-cluster sum of squares (WCSS). The algorithm is widely popular for both its simplicity and speeds, though at the cost of sensitivity to outliers and the need for a priori knowledge about the number of clusters needs to know to optimally partition a dataset (Kanungo et al, 2002). It is important to choose the correct number. Each cluster has a centroid, sometimes called a mean, a point from where the distance to all data points will be calculated. Hence the name “k-means”. K-means is called an iterative partitioning method as with each pass of the algorithm the mean value of the clusters is adjusted (Arimond and Elfessi, 2001). K-means follows a sequence of steps to identify a set of points as a cluster. K-means follows the following steps for each iteration. In the first iteration K data points are placed randomly in the data. During the second phase for each data point the closest cluster is determined. In the third phase, also known as the reduction phase the mean value of all points associated with the K data point is determined. In the fourth phase this mean is set as the new K data point. Steps 2 through 4 are then repeated until the K points no longer change or do so below a certain threshold.

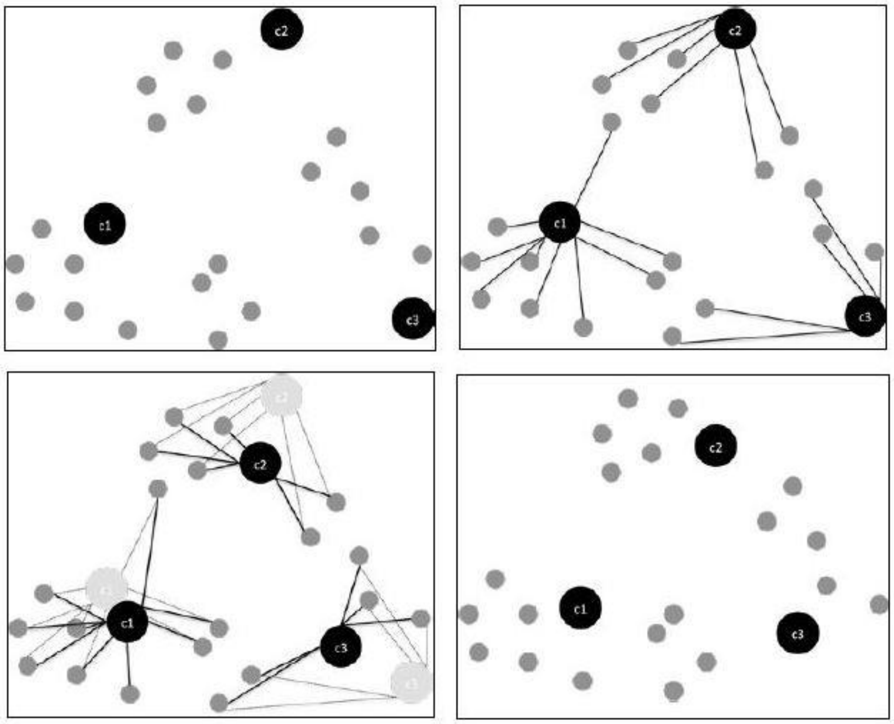


Figure 6 Steps taken by the K-means algorithm

Choosing the correct number of clusters K is mostly a matter of experience and evaluating results. However, there is a rule of thumb to determine the number of clusters you should initially investigate with a dataset. The number of clusters (k) is equal to the square root of the number of objects (rows) divided by two. In case of a data set of 200 rows, this will yield 7 clusters (Bell, 2014).

$$k = \sqrt[2]{\text{objects} / 2}$$

However, as datasets become larger so will the number of clusters according to this rule of thumb. In reality this is not case, rarely does K-means return more than 10 clusters. Bell (2014) suggests the elbow method as an alternative. The method relies on calculating the variance of the dataset as a percentage and plot this against the number of clusters. There will be an optimum number of clusters after which the increase of variance tapers off quickly. This point can be used to set as the number of clusters that the algorithm should try to discover. The elbow method is also used to determine the optimum number of variables with Principal Component Analysis (PCA). Despite the search for an optimal number of clusters the purpose of the clustering assignment should also be taken into consideration, the need for more detailed results at the cost of speed and generalizability will mean that there need to be more clusters (Pham et al, 2004)

Limitations of K-means

Besides the difficulty in determining the correct number of clusters that the algorithm needs to partition there are other drawbacks. K-means can be computationally intensive for large datasets. Per iteration the time it takes to calculate the mean is equal to product of the number of clusters and the number of patterns (Alsabti, Ranka and Singh, 1997).

Alternatives to K-means

One alternative to K-means is K-medoid, in which the objective is to minimize the Euclidean distance to the nearest center (Arora and Raghavan, 1998). K-medoid, also known as Partitioning Around Medoids or PAM, achieves better results than K-means when the data set contains noise and outliers.

K-means Extended or simply X-means uses the Bayesian Information Criterion (BIC-value) as part of an ‘Improve-Structure’ part. In essence X-means tackles each iteration as a 2-means problem. For each of the clusters a decision is made based on BIC whether it should be split further. X-means has the advantage in that it finds the optimal number of clusters on its own through analysis of the BIC-value.

Yet another alternative to K-means is K-median, which uses the median value for the data set across dimensions. Unlike K-medoid the median does not need to be an instance of the dataset. This factor increases flexibility as the initial k-median is not reliant on a data instance. The X-means alternative implementations of K-means will also be used and evaluated.

Hierarchical clustering

Hierarchical Clustering or HCA seeks to build a hierarchy of clusters. Such results are often displayed in a dendrogram or tree diagrams to show the relationship between data points as determined by the HCA algorithm. This method of clustering is divided into two strategies (Jesri et al, 2012). HCA is reliant on two concepts: the distance metric that is used to determine the distance between data points such as Euclidean or Manhattan distance (see section 5.6 on Normalization) and Linkage criteria to determine whether data points are similar.

- **Agglomerative.** Each data points initially forms its own cluster. Pairs of clusters are merged when the linkage criteria reaches a threshold.
- **Divisive.** All observations start in one cluster, and splits are performed recursively based on satisfying a distance parameter. If no stop rule is applied then by the end the number of clusters equals the number of data points.

Both types of strategies are known to be very computationally intensive. Hierarchical clustering tends to be more sensitive to data noise (Chaudhari and Parikh, 2012). HCA can use algorithms such as K-means to merge and split clusters. For this research an agglomerative algorithm is used which uses Ward’s method as the criterion to merge clusters at each step. Ward’s method is based on the Sum of Square Error (SSE) to determine minimal variance between data points (Hourdakis et al, 2010).

Density-based clustering

The Density-based clustering method relies on finding clusters based on the density of data points within a region. The number of clusters depends on whether each will have a minimum number of data points within a set radius determined to be the center of the cluster (Ester et al., 1996). A commonly used implementation of this type of clustering is DBSCAN proposed by Ester et. al. in 1996. Density-based algorithms have the ability to find any arbitrary shaped cluster with minimal interference from outliers. However, DBSCAN has difficulty discovering nested clusters. The alternative, OPTICS, does not have this deficiency but is very sensitive to fine-tuning of input parameters (Roy and Bhattacharyya, 2005).

Chaudhari and Parikh (2012) state that density based methods are not suitable for data with high variance in density. This problem occurs when there is for example two closely grouped clusters. Unlike other clustering algorithms such as K-means it is possible to find non-linear shaped clusters. An example would be a kidney shaped cluster next to a round cluster. This is possible as long the density of data points is maintained. If a dataset consists of a mixture of Gaussian distributions than density-based algorithms are regularly outperformed by Expectation–maximization algorithm. Variations, such as enDBSCAN, of the algorithm exist to solve this problem.

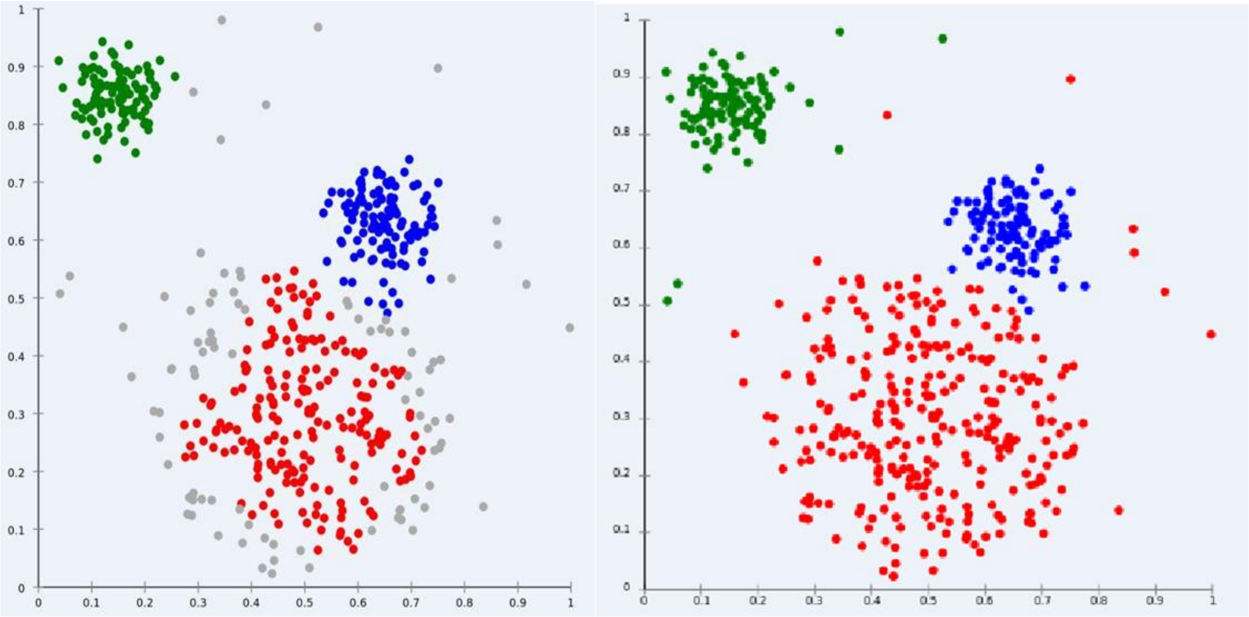


Figure 7 ordinary density vs. enDBSCAN

Distribution-based clustering

Expectation-Maximization (EM) algorithm.

The EM-algorithm has been used almost as long as the K-means algorithm. Despite being proposed it was not formalized until 1977 by Arthur Dempster, Nan Laird, and Donald Rubin. Unlike the K-means algorithm the EM algorithm is considered a soft clustering method. This method of clustering is based on distribution models. Unlike K-means each data point is assigned to a cluster that most likely has similar data points (Meila and Heckerman, 2013). The algorithm works by following an iterative process during which it calculates the membership probability for each data point under the given variables (Expectation step). In the second step (Maximization) this quantity is maximized (Dempster et al, 1977). An advantage of this approach is that is that data points can have multi-membership as they each have a certain probability to belong to a cluster. The EM algorithm is also able to deal with missing values better than most other algorithms. A downside to EM algorithm is that it can suffer from overfitting, it can also be complex to implement (Couvreur, 1997).

4.3. Semi-supervised clustering

A third variant of clustering is possible besides supervised and unsupervised learning. Semi-supervised learning small amounts of labeled data to aid in inferring a model or function. With supervised-learning the amount of labeled data is often limited. Semi-supervised learning circumvents the necessity for all data records to be labeled which can be resource intensive as it usually requires a skilled human agent or a physical experiment. It is also lessons risk of the results being an artifact of the algorithm. According to Basu et al. (2004) "unsupervised clustering can be significantly improved using supervision in the form of pairwise constraints, i.e., pairs of instances labeled as belonging to same or different clusters". Semi-supervised clustering falls info two general categories: constraint-based and distance-based.

4.3.1. Constraint-based methods

Constraint-based methods rely on user-provided labels or constraints to guide the algorithm towards a more appropriate data partitioning (Basu et al., 2004).

4.3.2. Distance-based methods

An existing clustering algorithm that uses a particular clustering distortion measure is employed: however, it is trained to satisfy the labels or constraints in the supervised data (Basu et al., 2004).
According to Basu et al. (2004) the use of constraint-base supervision is more general than the use of class-labels as a set of points that are classified imply a equivalent use of pairwise constraints, but not vice-versa. Their model is based on supervision provided in the form of must-link and cannot-link constraints, which indicates whether data points should be in the same cluster or not. This method also uses a penalty system where violations of constraints is penalized depending on the distance between data points. Closely lying cannot-link points are penalized more severely than those lying further away. Vice versa is true for must-link data points. Lange et al. (2005) agree that constraints can be particularly beneficial in data clustering where precise definitions of underlying clusters are absent. According to Basu et al. (2004) an important success factor of partitioned clustering algorithms such as K-Means is the choice of initial centroids. In their work on seeding such centroids they have shown that using labeled data points for limited supervision results in good initial centroids (Basu et al., 2002).

4.3.3. Implementation of constrained semi-supervised clustering algorithm

As the number of machine learning algorithms has reached many thousands I choose the implementation of Constrained K-means (COP-KMEANS) by Wagstaff et al. (2001) to explain the general method employed by semi-supervised algorithms. As the name of their algorithm suggests they adopted the very successful K-means algorithm to accept constraints such as 'must-link; which means data points must be within the same cluster and 'cannot-link' which means the opposite.
A generalized implementation of their COP-KMEANS algorithm looks as follows.

1. Let C_1 … C_k be the initial cluster centers.
2. For each point d_i in dataset D, assign it to the cluster C_j such that no 'must-link' or 'cannot-link' rule is violated. If no such cluster exists, fail return.
3. For each cluster C_i, update its center by averaging all of the points d_j that have been assigned to it.
4. Iterate between 2 and 3 until convergence.

5. Return {C_1 … C_k}

With this algorithm available background information can be used to guide the clustering to a favorable results. It also prevents data labelling from becoming an intensive and expensive undertaking. The results Wagstaff et al. achieved on a set of 6 diverse datasets are a marked increase in accuracy compared to just the use of K-Means or constraint rules.

4.4 Supervised Clustering

Supervised learning is based on the premise that values can be explained based on values of other variables, thus allowing groups to be discriminated. Common algorithms to grow such trees include C4.5 which is a Decision Tree algorithm, Random Tree and Bayesian Network algorithms. A common Decision Tree algorithm is CART, also known as 'Classification and Regression Trees' that first grows the tree to its full size and afterwards prunes the tree until the accuracy of the tree is similar for both the training dataset and the test dataset. Another method, CHAID or Chi-squared Automatic Interaction Detection' uses a statistical stopping rule to keep the tree from growing to impossible sizes. Decision trees can suffer from over-training, whereby the trees continue to grow and might afterwards not be able to validate test-data because it uses rules learned from the training data that are incompatible with the test data. Both CHAID and CART use different ways to limit the growth of decision trees (Caruana & Niculescu-Mizil, 2006).

Previous successful applications of supervised algorithms such as Decisions Tree, Random Tree and Neural Networks include handwriting recognition, image recognition and even whether an open source software project is successful or not (Amrit & Piggott, 2013). The latter research yielded a model of software projects could be classified into categories of development based on known metrics such as the number of developers, patches released and what operating system was supported. This decision model then becomes a predictive model for future projects. A similar application is envisioned for the KLM data after unsupervised clustering has been applied.

4.5 Clustering performance

According to Arimond and Elfessi (2001) a major challenge with performing segmentation research is finding a clustering method that can use qualitative (categorical survey) data.

4.6 Cluster Validity

Many times, cluster analysis is conducted as part of an exploratory data analysis. Hence, evaluation seems like unnecessarily complicated addition to what is supposed to be an informal process. Furthermore, since there are a number of different types of clusters – in some sense, each clustering algorithm defines its own type of cluster – it may seem that each situation might require a different evaluation measure. For instance, K-means clusters might be evaluated in terms of the Sum of Squared Error (SSE), but for density-based clusters, which need not be globular, SSE would not work well at all.

The problem with clustering is that almost every clustering algorithm will find clusters in a data set, even of that data set has no natural cluster structure. With high dimensional data such a scenario cannot be easily detected by visually checking the results. There are also different criteria for validity depending on the clustering techniques used. Compared to supervised learning procedures an unsupervised procedure is more difficult to assess as prior knowledge is not available.

Several questions need to be asked regarding the application of clustering methods.

1. Are there clusters in the data? This question can also be asked as 'Does the data set have a tendency to cluster data?'. Clusters are defined in this case as non-random structures in the data set.
2. Are the identified clusters in agreement with the prior knowledge of the problem?
3. Do the identified clusters fit the data well?
4. Are the results obtained by a method better than those obtained by another?

The first question can only be answered by trial and error, using several algorithms to determine whether there is any clustering tendency of the data. The other questions can only be answered after the application of clustering methods to the data. They form together the validation criteria. Jain et al (2010) distinguishes between three validation procedures.

- **External validation** consists of finding an answer to the second question and can only be performed when prior knowledge of the problem is available. Examples range from known general characteristics of the clusters and relations between specific items.
- **Internal validation** concerns the third question above and is based on an evaluation of the 'agreement' between the data and the partition.
- **Relative comparisons** attempt to provide an answer to the fourth question above and are usually the main application of the indices defined for the internal validation.

Various criteria according to clustering method

- **Unsupervised**. Measures of cluster validity are often divided into two class: measures of cluster cohesion (compactness, tightness), which determine how closely objects in a cluster are, and measures of cluster separation (isolation), which determine how distinct or well-separated a cluster is from other clusters. These are known as internal indices.
- **Supervised**. Clustering structures are compared with some external structure. These are known as external indices.
- **Relative**. Two cluster results are compared. With K-means the SSE value is a popular metric.

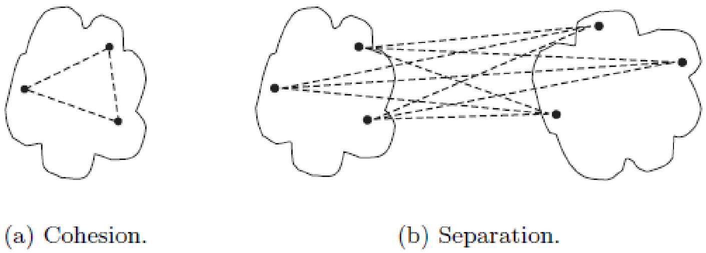
Unsupervised clustering validity.

Cluster cohesion and separation.

A recurring problem with applying clustering algorithms is that any data set will result in clusters. Even random data points will be clustered by algorithms such as K-means that partition using mean values distance values. As some points are closer to a randomly selected initializer while others are further away in the data set dimensional space clusters will form. Such clusters do not convey any specific knowledge to be interpreted. Two metrics that reflect whether clusters are more than random noise are cohesion, separation or some combination of these quantities. Cluster cohesion measures how closely related data points in a cluster are. Cluster separation measures how distinct or well-separated a cluster is from other clusters. Both can be calculated using the Sum of Squared Error (SSE). By comparing values within clusters to those between clusters cohesion and separation can show how well the dataset has the tendency to cluster. An alternative method based on proximity graphs calculates the sum of all weights within a cluster and between a node and all nodes outside a cluster.

$$cohesion(C_i) = \sum_{\substack{\mathbf{x} \in C_i \\ \mathbf{y} \in C_i}} proximity(\mathbf{x}, \mathbf{y})$$

$$separation(C_i, C_j) = \sum_{\substack{\mathbf{x} \in C_i \\ \mathbf{y} \in C_j}} proximity(\mathbf{x}, \mathbf{y})$$


(a) Cohesion.

(b) Separation.

Figure 8 Cohesion and separation visually depicted

A popular ensemble method calculating cluster cohesion and separation is the Silhouette Coefficient which combines both metrics for use on data points, clusters and clustering's. It is calculated as follows.

- **Step 1:** calculate value A of a point i_1 from a cluster the average distance to all other points within that cluster.
- **Step 2:** calculate value B the average distance between point i_1 and the same number of other points found in other clusters.
- **Step 3:** in the final step the Silhouette Coefficient is calculated by dividing the value A by B and subtracting the result from 1. If the distance between i_1 and those points in another cluster is large than value B will be much larger than A and A/B will tend towards 0. Subtracting this from 1 means for a Silhouette Coefficient between 0 and 1 the latter is more desirable.

The objective of the Silhouette Coefficient is to assign one metric that would prove a cluster solution has both high intra-cluster similarity and low inter-cluster similarity. This is an *internal criterion* for the quality of a clustering.

One can use Hopkins statistics to see whether data will cluster well. Values at 0 or 1 are good while close to 0.5 are bad. For hierarchical clustering you have to use the cophenetic distance, which is the proximity at which an agglomerative hierarchical clustering technique puts the objects in the same cluster for the first time. It can be considered a measure of how faithfully a dendrogram preserves the pairwise distances between the original un-modeled data points.

Number of clusters

There are several determinants to judge what the optimal number of clusters may be. One method reminiscent of Principal Component Analysis. To determine the number of clusters a SSE graph is used. This technique is made difficult when cluster are intertwined or overlapping.

Supervised and Semi-supervised clustering.

For supervised clustering metrics such as purity, precisions, recall and F-measure can be used. These can also be used for validation with unsupervised clustering if the results are considered prior knowledge.

Precision.

The fraction of a cluster that consists of objects of a specified class.

$$\text{P} = \frac{correctly\ assigned\ to\ cluster\ A}{Assigned\ to\ cluster\ A}$$

If 100 data points are assigned to cluster A and 80 are actually correctly assigned to A than the ‘Precision’ has been 8/10 or 80 %. This metric does not take into account data points that should have been assigned to cluster A. A high precisions has meant that an algorithm has identified correctly substantially more data points as part of a cluster than not.

Recall.

The extent to which a cluster contains all objects of a specified class.

$$\text{R} = \frac{correctly\ assigned\ to\ A}{labeled\ as\ A}$$

If 80 data points are correctly assigned to cluster A out of 120 that are actually labeled to be part of A than ‘Recall’ has been 8/12 or 66 %. A high recall means that most of the labeled data points are assigned to the correct cluster.

F-measure.

A combination of both precisions and recall. It calculates the harmonic means between both indices. A value of 1 is considered best and 0 is worst.

$$\text{F-measure} = \frac{precision * \ recall}{precisions + recall}$$

Cluster purity can be calculated as by taking the number of correctly clustered data points divided by the total number of data points. This method however cannot say anything about the number of clusters not their quality (Manning et al. 2002).

The Normalized Mutual Information says something about the cluster count and their quality. One can even compare results of clustering which have different amount of clusters.

Rand index

A very important statistical metric for external validity is clusters is the Rand-index. This measures the percentage of objects that are clustered correctly. It is calculated by dividing the total number of correct data points clustered by the total number of data points in a data set. It is based on the following indices.

1. True positive (TP). Data points correctly clustered.
2. True negative (TN).Data points correctly assigned to another cluster.
3. False positive (FP). Data points incorrectly assigned to a cluster.
4. False negative (FN). Data points incorrectly assigned to another cluster.

As labeled data may contain more than 1 category all ‘True Negative’ data points can be considered as ‘True Positive’. All other data points can be grouped together. This will yield the Rand index for the entire cluster assignment.

Prior knowledge

A unique case for checking cluster validity exists after the application of unsupervised clustering algorithms. When data points have been clustered can be labeled as belonging to a cluster. Afterwards supervised learning algorithms such as Random Tree and Decisions Tree can be applied to validate clustering results based on metrics such as Precisions, Recall and F-measure. Such validation methods are little referenced in literature but are widely used in practice (Kishida, 2014). To confirm the results a random clustering result on which supervised algorithms are applied is used to check the method applies to a dataset. In theory, the F-measure should be 1 divided by the number of clusters. The main advantage of using supervised learning algorithms for validation is that it can used across all unsupervised learning results.

4.7 Normalization

The following sections discuss various topics that need to be taken into consideration when using Machine Learning algorithms. With Machine Learning algorithms such as Neural Networks the neurons that are used give outputs of 0 and 1. If the target values are not 0 or 1 than they should be scaled so that they are. This helps prevent the weights from becoming too large unnecessarily (Marsland , 2014). Another way to prevent this is to scale the inputs. A common method is to treat each dimension in the dataset separately and scale them to ensure that the minimum value is -1 and the maximum is 1.
The advantage of feature scaling is two-fold.

1. The first advantage has to do with calculating the distance between two points. If one feature has a broad range of values (beyond the scale of -1 and 1) while other features do not than this feature will govern the distance.
2. The second advantage for feature scaling is that algorithms such as gradient descent converge much faster with feature scaling than without it.

The general formula for feature scaling is given as: X’ $\frac{x-\min(x)}{\max(x)-\min(x)}$

Here X’ is the scaled value and X the original value. As an example, consider we want to cluster passengers that are eligible for enrollment in Flying Blue tiers Silver and higher. The minimum number of required flights is 15 while one person has performed many as 141 qualified flights. To rescale these values we first subtract minimum value 15 from each passengers qualified number of flights and divide that number that number by 126. All values will be scaled to between 0 and 1.

A closely related topic is Multi-Dimensional Scaling (MDS). This allows various variables to be plotted together for easy comparison. It is considered an alternative to Factor Analysis. As in factor analysis, the actual orientation of axes in the final solution is arbitrary. MDS is not so much an exact procedure as rather a way to "rearrange" objects in an efficient manner, so as to arrive at a configuration that best approximates the observed distances. It actually moves objects around in the space defined by the requested number of dimensions, and checks how well the distances between objects can be reproduced by the new configuration.

Normalization is considered to be somewhat of a black art in that many of the procedures that should take place during pre-processing are not well defined. For use in K-Means a normalized dataset can give completely different results. This is due to changes in the Euclidean distance between data points, K-Means is highly dependent on finding the nearest neighbor in the Euclidian space defined by the data.

4.8 Curse of Dimensionality

The curse of dimensionality refers to the problem that occurs when the number of input dimensions grow there also needs to be more data points. Training classifiers will then take longer.

The problem can be visualized with a hypersphere. In essence the curse of dimensionality can be stated that as the number of dimensions increase the volume of the hypersphere does not increase with. Instead it tends to zero. In a two dimensional space drawing all points that are at a distance of 1 from the origin creates a circle. In a three dimensional space it is a sphere around origin 0.0.0. Notice how for this example the sphere takes up relatively less space of the three dimensional cube than the circle would in the two dimensional space defined by the rectangle.

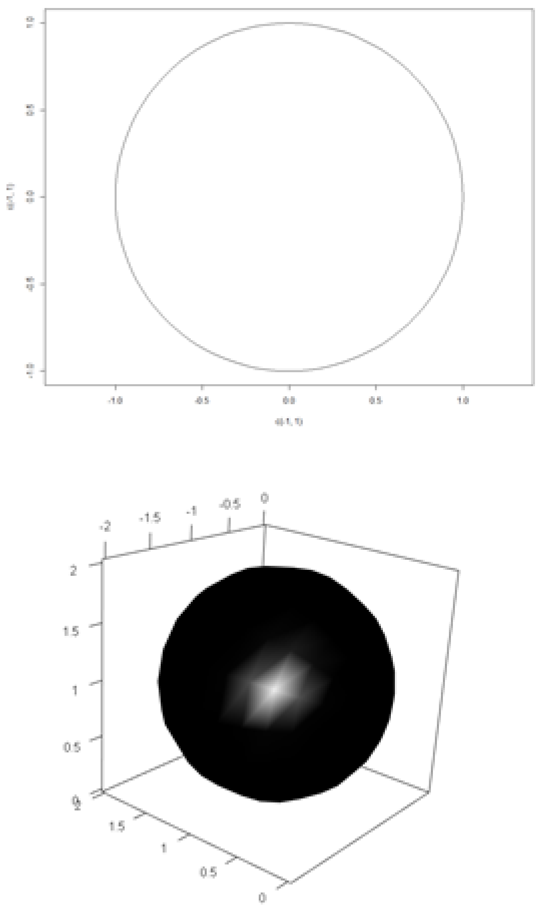


Figure 9 Reduction of space as dimensions increase

The reduction of the volume of the hypersphere is directly related to the ability of Machine Learning algorithms to generalize sufficiently well. With a fixed number of training samples, the predictive power reduces as the dimensionality increases, and this is known as the Hughes effect or Hughes phenomenon.

One way to deal with the Curse of dimensionality is to perform dimensionality reduction. This produces lower dimensional representations of the data that still include the relevant information. The goal of reducing data dimensions is to uncover data dimensions that will still allow to separate out different classes. Intrinsic Dimensions refers describes how many variables are needed to represent the variable.

However, some algorithms that are based on distance functions or nearest neighbor search can also work robustly on data having many spurious dimensions, depending on the statistics of those dimensions.

Choosing the number of variables

Principal Component Analysis or PCA results in components that account for a maximum amount of variance for observed variables. There are several methods with which dimensions can be reduced. The most common is PCA. With PCA the algorithm first centers the data and then places an axis along the direction with the largest variation. It then places a second axis that is orthogonal (perpendicular) to the first and that will cover as much of the remaining variation as possible. This process is continued until it runs out of possible axis. The end results is that all the variation is along the along the axes of the coordinate set, and each new variable is uncorrelated with every variable except itself. Those axes that show very little variation can be removed without affecting the variability of the data.

Principal Component Analysis is based on Linear Algebra, the axes that are placed along the direction of the greatest variation are essentially the Eigenvectors of the covariance matrix. PCA is sensitive to the relative scaling of variables. This makes sense if you consider where the axes are placed. To prevent PCA from simply labelling variables in order of scale they will need to be normalized before PCA is applied.

With a normalized dataset the criteria for whether a dimension should remain is the Eigenvalue associated with an Eigenvector. If these values are close together the dataset is already in a ‘good’ subspace. If some values are higher than others then consideration should be given to only keeping those with a high value. Dimensions with Eigenvalues close to 1 or 0 have no descriptive value (Raschka, 2014).

There are several steps that need to be taken before clustering algorithms can be applied to a data set. After the data set has been built the first question that needs to be answered is “How many dimensions fit the given data?”. To determine how a given configuration (n-points in a t-dimensional space) fits the data a stress measurement will be used. A popular stress test was defined by Kruskal (1964) who stated stress to be: “a residual sum of squares, it is positive and the smaller the better”. The objective then becomes to discover as many and which variables within the dataset that can quickly reduce the residual sum of squares (residual variance) that still exists. Kruskal concluded the following table to be a good indicator of when the configuration of the data set will fit the underlying data.

Stress	Assessment of fit
20 %	Poor
10 %	Fair
5 %	Good
2.5 %	Excellent
0 %	“Perfect”

Table 2 Kruskal (1964) data set configuration fit

With a scree plot, the stress assessment can judged along with the number of dimensions in the data set. The interesting point in the plot is where the addition of more dimensions does not significantly reduce the residual sum of squares. This method is known as the ‘elbow-method”. It may be considered subjective but it is considered effective. This approach will also be used for this research to determine the best configuration. SPSS is used to perform a factor analysis. This includes a scree plot which contains the eigenvalue for components with which the ‘elbow-method’ can be used. It also has a table with the total variance explained which contains a column with cumulative variance of all preceding components. This can be used to determine how many components are need to achieve the data configuration fit described by Kruskal (1964). Finally a component matrix show which variables have the greatest influence on the components of the scree plot and the table with the total variance. This can be used to determine which variables are suitable to select for a configuration that is to be used for cluster analysis.

4.9 Regularization

One method commonly applied to Machine Learning algorithms to prevent overfitting is Regularization. By using Regularization a form of feature selection is immediately applied. It can be considered as an alternative to cross-validation which focuses on repeatedly performing similar uses of an algorithm to check consistency. Regularization usually consists of introducing a penalty for additional complexity to ensure the model the algorithm extracts is not more complex that it needs to be. An example already discussed is BIC or Bayesian Information Criterion.

5. Methodology

5.1. Data set analysis

This chapter will give an overview of the steps that were taken to complete this research. As far as possible technical details are avoided or are mentioned in other chapters together with the theory on machine learning algorithms and result validation. The steps that were taken are framed into context with the CRISP-DM methodology which is used to facilitate Data Mining projects. CRISP-DM was used as a general guideline to perform this research effort in structured manner with the hopes that the process is reproducible to answer future business questions. As this research is a data mining effort using data clustering algorithms it could not be guaranteed that the results would answer the research questions. As such the research effort should be conducted in a flexible manner.

According to Dolnicar (2002): “The basic idea of cluster analysis is to divide a number of cases (usually respondents) into subgroups according to a pre-specified criterion (e.g., minimal variance within each resulting cluster) which is assumed to reflect the similarity of individuals within the subgroups and the dissimilarity between them” (p. 4). There are however a number of analysis steps that can be taken to discover whether a data set can possibly yield a successful cluster analysis.

1. The first step would be to perform a descriptive analysis of the dataset. The objective is to become familiar with the dataset and determine what its potential is.
2. The second phase will be an exploratory analysis. The objective is to find relationships (correlations) in the data that were not previously known in order to determine whether the research questions that have been asked will be feasible. Such an analysis can give an indications whether the dataset will eventually yield clusters.
3. The final step can either be a inferential or predictive analysis. The goal of the former is to use a small portion of the data to say something about the larger population. Predictive analysis is essentially models the whole population so we can identify in which cluster a customer can be segmented as soon as they are passengers with KLM.

The results of these steps can be read in the chapter describing the dataset

5.2. What is CRISP-DM?

When Data Mining showed signs of exploding into widespread uptake in the 1990’s employees of DaimlerChrysler questioned whether their approach was the right way. They had learned their skill by trial and error and wondered whether other early adopters should have to go through the same process. To show the value and maturity of Data Mining a standard process model was devised by a consortium.

CRISP-DM is intended to be industry-, tool- and application neutral hence the acronym which stands for (Cross-Industry Standard Process for Data Mining). CRISP-DM has not been built in a theoretical, academic manner working from technical principles. It is instead based on practical, real-world experience of how people conduct data mining projects (Chapman et al, 2000).

5.2.1. Hierarchical breakdown

CRISP-DM is a process model of a set of tasks described on four levels of abstraction (from general to specific): phase, generic task, specialized task and process instance (Behja et al, 2012). Both the first and second level are intended to describe generic processes that cover all possible data mining applications. The third level, specialized task, would describe how generic tasks should be carried out. As an example it should describe whether a dataset should have either the numerical values or categorical values cleaned as part of the generic task of data cleaning. The fourth and final level, the process instance, is a record of the actions, decisions and results of an actual data mining engagement.

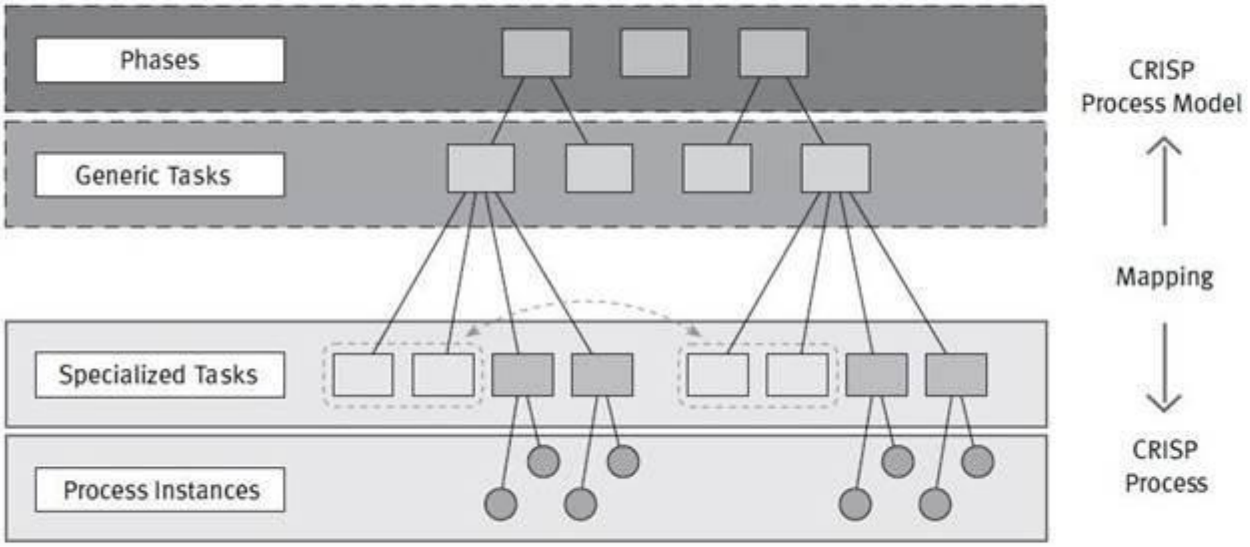


Figure 10 Process model hierarchy from CRISP-DM 1.0 Step-by-step data mining guide (Chapman et al, 2000)

5.3. Further CRISP-DM literature

As CRISP-DM has been used for almost two decades a large amount of literature describes its practical use. What follows is a selection of the most pertinent papers.

According to Alsultanny (2011) “Schumann (2005) proved that the CRoss-Industry Standard Process for Data Mining (CRISP-DM) can be transferred to an educational settings and provide a start-to-end structure that is capable of producing operationally actionable information.” The author concludes that “CRISP-DM is a non-proprietary data mining process that was developed for and is currently used in the business world. This proves that the CRISP-DM method has the possibility of being generalized and be widely applicable.”

Nadali et al. (2011) have investigated the success levels of data mining projects that are based on the CRISP-DM method. The failure rate of Data Mining projects may actually be as high as 60%. Despite the prevalence of methodologies most data mining projects are still performed in an unstructured and ad hoc manner (Becker & Ghedini, 2005. According to the authors Nadali, Kakhky and Nosratabadi (2001) “The successful conclusion of each phase of the CRISP-DM method will be important for the success of the subsequent phase and the overall project. Adequate phase evaluation can thus improve the success of Data mining projects which further strengthens the merits of having an industry wide standard process model".

Sharma et al (2012) describe KDDM process models as follows: "Knowledge Discovery and Data Mining or KDDM process models serve the purpose of a roadmap or guide, that provide prescriptive guidance towards how each task in the end-to-end process can be implemented. They can be regarded as a reference guide or manual that describes 'what' tasks should be executed in the context of a Data Mining project and 'how' they should be executed."

5.4. How can CRISP-DM be applied?

The CRISP-DM process model provides an overview of the life cycle of a data mining project. It describes the phases of the project, their respective tasks and the relationships between these tasks. CRISP-DM has six phases a shown in Figure 10. Their sequence is not rigid. In fact, moving back and forth between phases is required. The arrows in the figure show the most important and frequent dependencies between phases. The time it takes to perform a cycle is also not rigid. Data mining projects have their goals frequently adjusted which requires new iterations through the CRISP-DM life cycle. Data Mining usually does not end with a solution, instead they trigger new projects often with more focused business questions. Below a overview of each step in the life cycle is given along with a brief description of relevant issues encountered during this research effort

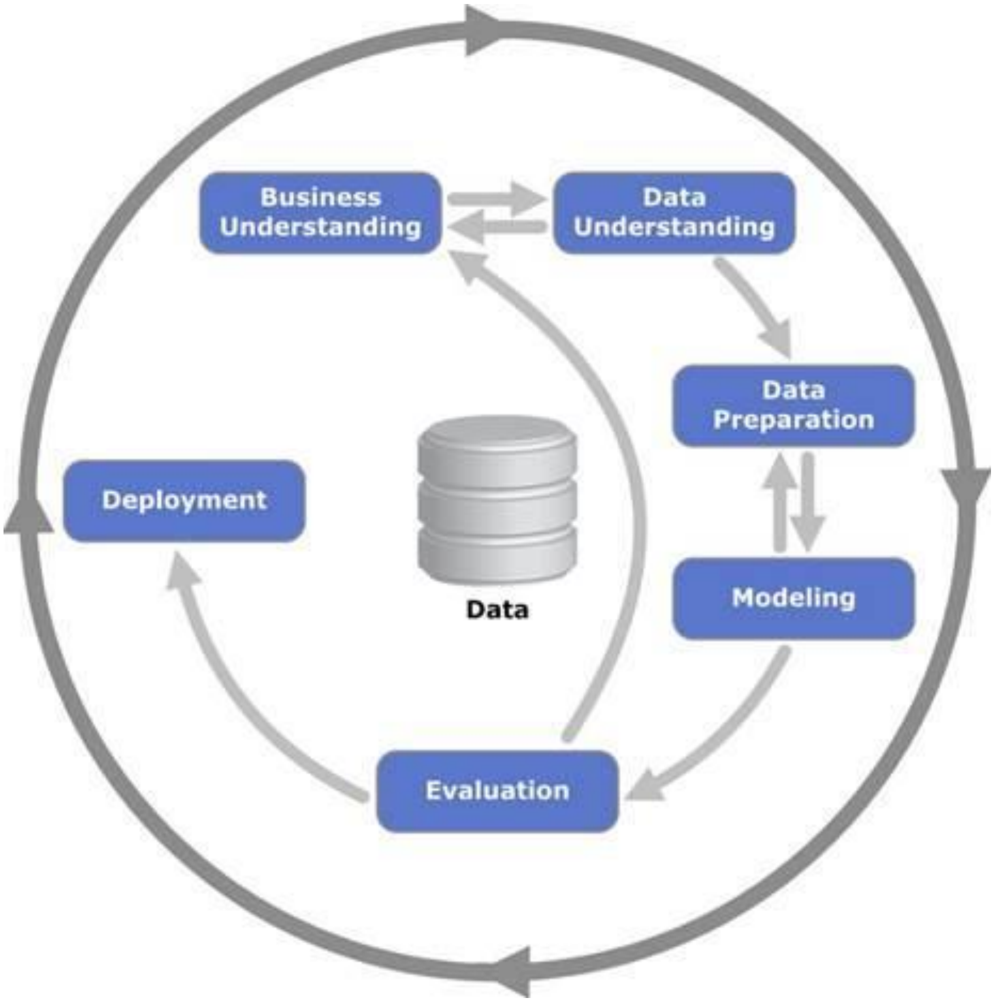


Figure 11 Life cycle from CRISP-DM 1.0 Step-by-step data mining guide (Chapman et al, 2000)

5.4.1. Business understanding

This initial phase focuses on understanding the project objectives and requirements from a business perspective. This knowledge is converted into a data mining problem definition and a preliminary plan is designed to achieve the objectives.

In this preliminary phase the goals of the stakeholders need to be matched with the availability of the data set and analysis tools. A more formal stakeholder analysis was conducted after one iteration of the CRISP-DM cycle (lasting a few weeks) to delineate the scope of the research effort. Once the research effort yielded clustering results that were usable for the purposes of market segmentation the deliverables were decided upon with stakeholder to ensure results would be of use for the business.

5.4.2. Data understanding

The data understanding phase starts with initial data collection and proceeds with activities that enable you to become familiar with the data, identify data quality problems, discover first insights into the data, and/or detect interesting subsets to form hypotheses regarding hidden information.

As the principal data set used had 77 variables a choice has to be made as to which can be used to cluster passengers into market segments. Variables such as distance traveled and frequency of travel could foreseeably be aggregated into a new variable average distance traveled. Previous market research was used to choose which data sets and which variables would be used to create the data set for the analysis. As some data sets only become available later in the research effort: steps 2 to 5 (Data Understanding to Evaluation) are repeatedly carried out over the course of several weeks.

5.4.3. Data Preparation

The data preparation phase covers all activities needed to construct the final dataset (data that will be fed into the modeling tools) from the initial raw data. Data preparation is likely performed multiple times and not in any prescribed order.

Cleaning data was the most common task performed within 'Data Preparation'. Missing values had to be interpreted: they can either mean a value 0 or incorrect data retrieval. To create an unique record for each airline passenger a key to distinguish passengers had to be discovered that would allow for a practical implementation of the final data set. As the key also consisted of missing values and noise a process of testing was necessary to determine if the flaws were random or statistically irrelevant.

5.4.4. Modeling

In this phase, various modeling techniques are selected and applied, and their parameters are calibrated to optimal values.

Modeling is lengthy process whereby a sample of the dataset is subjected to many clustering algorithms. During this process it will become clear which variables have the strongest descriptive capabilities. As variables are removed and more are added this phase has a strong feedback relationship with the data preparation phase. To perform all data preparation before modeling would be a mistake. A partial result from modeling can already be evaluated which could alter the Data Mining goals.

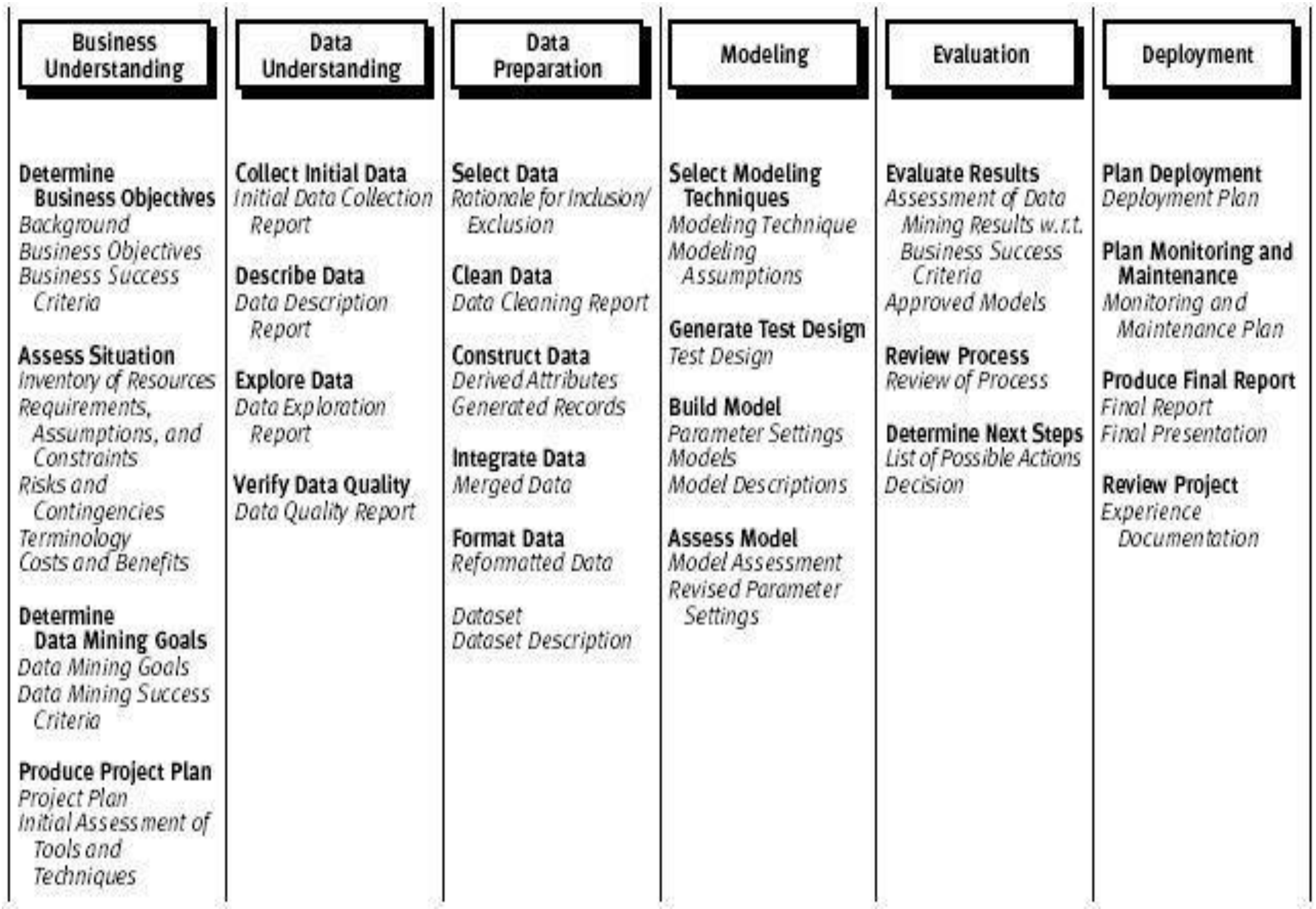
5.4.5. Evaluation

The purpose of this phase is to perform a thorough evaluation of the results from the modeling. A key objective is to determine if there is some important business issue that has not been sufficiently considered.

Clustering results can be considered too abstract for business understanding. For the purpose of increasing understanding revenue estimates were added to the dataset. For each sub class that a passenger flew with a revenue estimate was available. All the numbers were summed for each passenger. Not only did this improve business understanding and clarify relevance it also added an important variable for the purposes of data clustering.

5.4.6. Deployment

This phase can include just generating a report or implement a repeatable data mining process across an organization. For business stakeholders it is important to understand what actions need to be carried out in order to actually make use of the created models. For each of the six phases there are a large number of generic tasks that can be performed. An overview of each step is given on the following page.



Business Understanding

Determine Business Objectives
Background
Business Objectives
Business Success Criteria

Assess Situation
Inventory of Resources
Requirements, Assumptions, and Constraints
Risks and Contingencies
Terminology
Costs and Benefits

Determine Data Mining Goals
Data Mining Goals
Data Mining Success Criteria

Produce Project Plan
Project Plan
Initial Assessment of Tools and Techniques

Data Understanding

Collect Initial Data
Initial Data Collection Report

Describe Data
Data Description Report

Explore Data
Data Exploration Report

Verify Data Quality
Data Quality Report

Data Preparation

Select Data
Rationale for Inclusion/ Exclusion

Clean Data
Data Cleaning Report

Construct Data
Derived Attributes
Generated Records

Integrate Data
Merged Data

Format Data
Reformatted Data

Dataset
Dataset Description

Modeling

Select Modeling Techniques
Modeling Technique
Modeling Assumptions

Generate Test Design
Test Design

Build Model
Parameter Settings
Models
Model Descriptions

Assess Model
Model Assessment
Revised Parameter Settings

Evaluation

Evaluate Results
Assessment of Data Mining Results w.r.t. Business Success Criteria
Approved Models

Review Process
Review of Process

Determine Next Steps
List of Possible Actions
Decision

Deployment

Plan Deployment
Deployment Plan

Plan Monitoring and Maintenance
Monitoring and Maintenance Plan

Produce Final Report
Final Report
Final Presentation

Review Project
Experience Documentation

Figure 12 CRISP-DM generic tasks from CRISP-DM 1.0 Step-by-step data mining guide (Chapman et al. 2000)

5.5. Best practices

The following list of best practices was accumulated through the Structured Literature Review of this master thesis and can be considered guidelines.

1. Keeping the size of the data and variables manageable. According to Fayyad (1996) a dataset of a few thousand observations with just 15 to 20 variables is preferable. This is guideline is closely related to the 'Curse of Dimensionality' whereby each variable added to the dataset reduced descriptive power of the dataset. Where possible variables were condensed into aggregates of two or more variables (such as average distance)
2. The measure of association underlying the clustering algorithm is applicable to the data format. Data can consist of the following numeric type.
    a. **Ordinal:** values exist on a scale and have a clear ordering. Examples include questionnaire answers where a selection can be made from 1 through 5 to state your approval of a proposition, but A through E character scale is also possible if the scale and ordering remains clear.
    b. **Nominal:** this is similar to 'categorical' data, but there is no intrinsic ordering associated. Examples include gender (male or female) or hair-color. The variables have no special meaning vis-a-vis each other.
    c. **Metric:** values lie on an interval scale. Example include weight, distance and revenue.

   According to a study conducted by Dolnicar (2002) most use ordinal data (66%) followed by nominal data (23%) while the use of metric data is negligible. Yet others use a combination of ordinal and nominal data. Ordinal data is preferred. However, the airline dataset consisted entirely of Nominal (gender, corporate flag) and Metric data (age, frequency, distance traveled, weight and number of bags carried etc.).
3. Data pre-processing must be performed. This includes procedures such as factor analysis that reduces the number of variables by searching for underlying factors. However, factor analysis can be a double-edged sword. While it may remove superfluous variables and thus reduce compute cycles it can also ruin the use of some clustering algorithms that rely on the dependence between variables that should be mirrored in the clusters. Factor analysis with ordinal data is usually not necessary (Dolnicar, 2002). Factor analysis was used to end the process of adding more variables to the data set (see reference to chapter). Through factor analysis it is discovered that only about one third of the variables are highly descriptive (they explain most of the variance) while the other two thirds are at best nominally descriptive.
4. Validation of results. If an external information source is available then content validity can be evaluated easily. Otherwise only 55% of studies perform validation with statistical measures and discriminant analysis being the most popular (Dolnicar, 2002). This research effort performed both validity through evaluations with external information sources as well as statistical measures. External sources indicated whether the clustering results coincided with possible market segments while statistical measures indicated whether the data set was adequately clean and contained the right variables. Both types of validation indicated when it was time to move towards the deployment of results.

6. KLM dataset

The dataset that is used in the attempt to cluster passengers is provided by KLM.

Introduction

This section describes the procedures that have been applied to the KLM datasets. The purpose of the procedures is to collate and clean-up the data to make it suitable for data analysis. The primary data is extracted from Altea, a system that records all passenger flight movements. This data is available in the form of flat files called DECODE, but they can also be accessed through a front-end system called Opera. However, Opera has limits to the size of the calculations that can be made and files that can be extracted. Other systems that were use include PSQN, which is solely booking data that feeds into DeLorean, a booking analysis tool and the Monet revenue system. On the next page schematic gives an overview of how these system are interconnected and how they were siphoned for this research effort.

Altea

Altea is a relatively new system that replaces Corda and Codeco. The process of replacing that system with Altea is ongoing but the primary system was completed in 2013 which makes 2014 the first year for which a complete record exist. For 2014 Altea consists of 12 files, one for each month, with a total of 19.246.730 records and a total size of 7.88 Gigabytes. Each record represents one passenger flight movement. Nominally with each flight movement there are some 77 associated variables ranging from flight details (airline, flight number, departure, arrival and date), personal details (first name, surname, birthday, gender) as well many variables about seating, frequent flyer programs, check-in method, ancillaries and information about baggage.

For the purpose of performing analysis on passenger's types the Altea dataset was altered so that for each unique passenger there exists only one record for 2014. The total number of unique records is 9.021.245. Thus each record represents a longitudinal record of the actions a passenger took over the course of 2014. The following paragraphs describe the actions that were taken to collate and clean up individual variables and the reason for why they are included for the purpose of analysis.

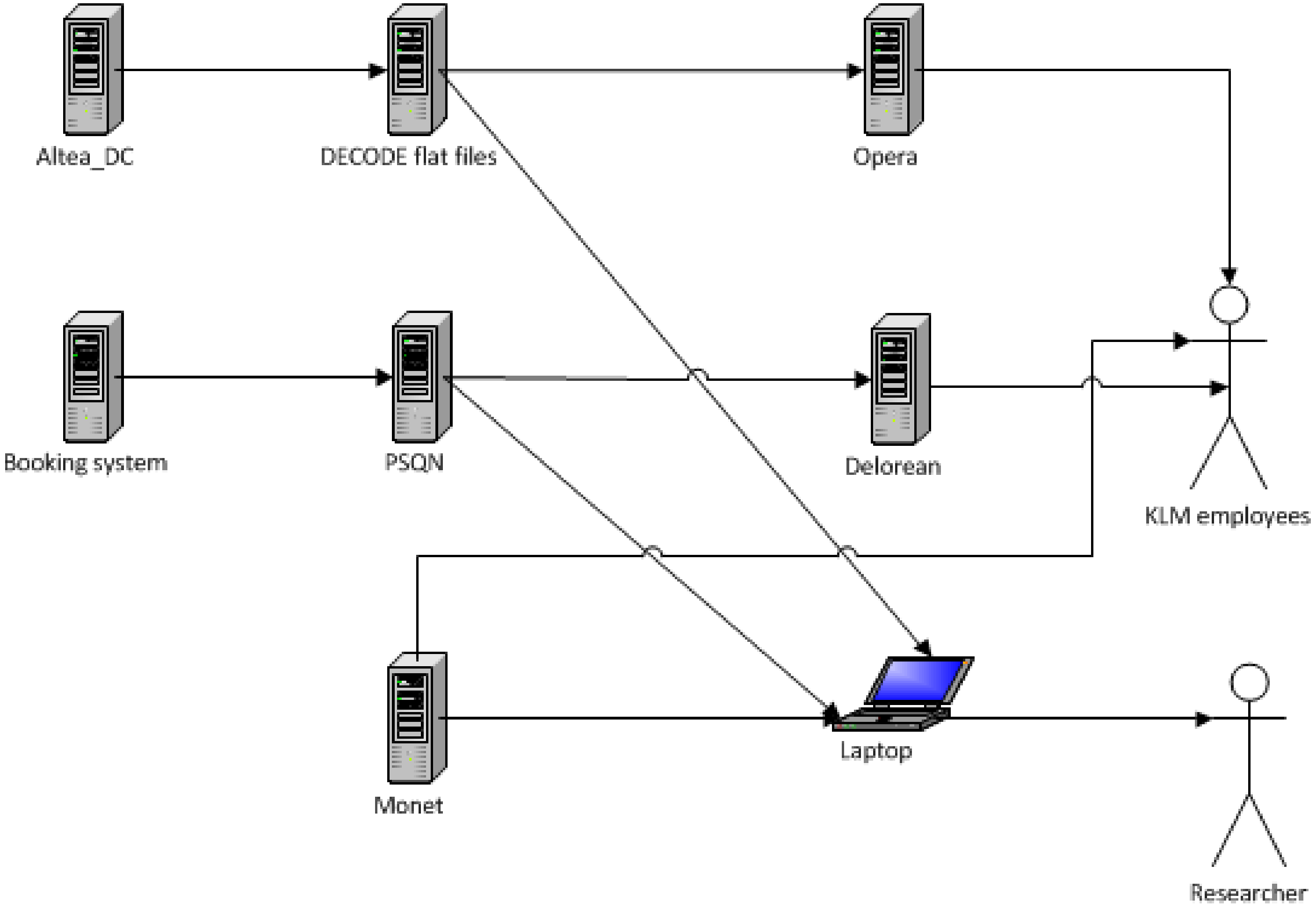


Figure 13 Data set overview

Primary Key

One of the most difficult decisions that had to be made was deciding on a key to identify passengers over a period of time. Those passenger not enrolled on any KLM frequent flyer program or contract are not obligated to declare all personal details. Furthermore no unique identifier is maintained by KLM. A key based on {surname, first name} is considered insufficient as records belonging to separate unique persons would become bundled together if they have a common first and last name. The addition of email address is problematic as passengers often book flights using different email addresses depending on the purpose of the travel. Altea also does not have email addresses stored, these could only be obtained through merging records from another dataset. The only alternative key is {surname, first name, DoB} which makes it unlikely that passenger records are bundled but has the added problem that not all records are complete. Sadly out of 9.021.245 records 3.300.449 have no DoB. Whether or not passengers are obligated to enter in their DoB is highly dependent on the travel destination. Such information as well as full name and gender are collected by KLM into an API (Advanced Passenger Information) and sent to the authorities of a country as required. However, not all countries require a full API. Besides this issue passengers also frequently fill in their DoB incorrectly. Nonetheless after comparing samples of the populations with and without a DoB the decision was made to proceed with this key.

Frequency

An important indicator for passenger motivation is the number of times a passenger flies within a period of time. Using the primary key {surname, first name, date of birth} each occurrence over the course of 2014 was counted.

A sample of the result shows that several passengers with very common first and last names have traveled around 200 times during 2014. There is no way to proof the records represent unique persons or several individuals. The first record with a perfect primary key that represents a passenger has traveled 141 times. The top of the data sample shows a large number of records without a birthday. This is due to the fact that people will share common names and counting their flights aggregates them at the top. After the initial 30 records the date of birth reaches a ratio around similar to the rest of the data.

Distance

Another critical variable that may indicate whether KLM passenger travel for business or leisure is the distance flown within a period of time. As Altea contains the departure and arrival codes for each flight it is possible with a table containing distances that correspond to those codes to calculate the distance flown per passenger. One import consideration remains with distance. If a person travels 5 to 10 times per year on business to a destination close to the point of departure then he may not be distinguishable from a leisure traveler who travels once or twice a year to a faraway holiday destination using just this metric. As such, the alternative metric 'Average Distance' has been added. For each passenger flight movement the distance was added, using the primary key all distances were summed for each passenger. There are 448 different airport combinations that KLM and subsidiaries have flown to over the course of 2014.

Model Redacted

Gender

According to the model of passengers based created by Arwed Wegscheid and Julia Godet those passengers identified as business traveler are predominantly male. As leisure travelers are likely to travel in pairs or larger groups the ratio of men and women should be more balanced. Altea data includes gender for each flight movement. No major inconsistencies were discovered. Gender is represented by just one variable with values 0 and 1 representing female and male respectively.

Age

Age may be another significant factor to identify different passenger types. Description Redacted For each passenger flight movement a date of birth is included if the Advanced Passenger Information request required it. The date of birth was then converted to age in years. The decision was made not to use incremental categories for age but instead keep it as a numerical value in order to better discriminate on this variable. Due to the fact that one third of passengers do not have a known date of birth it is vital to analyze the distribution of those that do. Below is a histogram of the age distribution. A major anomaly are those aged 95 and above, but they are statistically an insignificant number. The records with unrealistic ages are so small they are represented in the histogram merely by a line instead of a rectangle. Observations made of the rest of the dataset show that people do frequently make typos with their birthday. One frequent flyer has as a birth year 1982 and traveled 16 times. Under the same name and birth date, but with a different birth year he has also traveled 6 more times. The same name without a birthday again has traveled 6 more times as well. While it is virtually certain the first two records depict the same person the same cannot be said for the last record.

Histogram Redacted

Baggage amount and weight

Both these variables represent the sum value for each passenger over 2014. The model of business travelers predicts they will only carry hand luggage on small and medium haul flights while they carry one bag on long-haul flights. However, carry-on baggage is not recorded. One passenger has checked in over 4000 kilograms of luggage during 27 flights. The records do not show whether he was traveling alone.

Check-in method

This variable describes through what method passenger choose to check in at the terminal. Options include Internet, Manual, Kiosk or through an External Departure Control System (DCS). As passengers can alter their behavior the decision was made to depict each method as a variable and count each occurrence per passenger. An alternative variable included recorded the most popular choice of check-in method for each passenger. However, this leads to a problem with clustering algorithms that prefer numerical data as opposed to categorical data.

Example of output with fictional data:

	Check-in Method				
Primary Key	**Internet**	**Manual**	**Kiosk**	**External DCS**	**Most frequent**
Piggott James + DoB	5	1	1	1	Internet
Greeve Fai + DoB	4	1	0	0	Internet
Van Keulen Maurice + DoB	2	4	1	0	Manual
Amrit Chintan + DoB	0	0	2	1	Kiosk

Travel class

A numerical value is given for each time a passenger either flies Business, Economy Comfort or Economy class. Passenger behavior can change due to the purpose of the flight so this is reflected in the data. An added variable shows what the most popular travel class was for each passenger. Business travelers are predicted to travel more frequently in Business class or Economy Comfort than Economy class.

Frequent Flyer program

This is a Boolean variable that describes whether a passenger participates in a frequent flyer program (from KLM, Air France or Delta). Three more Boolean variables show which airline they participate in.

Tier Level

For each tier level of Flying Blue, the frequent flyer program, a counter is kept per passenger. As passenger movie up in the tier structure their progress can be seen. Options for tier level include: None, Ivory, Silver, Gold and Platinum.

Corporate flag

This attribute is depicted with Boolean value that indicates whether a passenger is traveling using a corporate account. Such accounts are suspected to be a good indicator for business travel. If a cluster has proportionally more members with a corporate account it may be that the other members are also business travelers who don’t have an account.

PSQN data

Another data set that is closely related to Altea is PSQN. This system uses booking data, not actual flight movement data. It subsequently feeds into DeLorean, the KLM booking analysis interface. PSQN contains data regarding the length of stay for each passenger at their destination, the point of sale of the ticket as well as the lead time between the booking and the flight. The latter may be calculated from Altea data but PSQN had the data available in an easy to use format.

Point of Sale

The point of sale of a booking represents the country from which the booking originates. In the KLM dataset countries are represented through 2 letter abbreviations. Commonly used abbreviations include NL (Netherlands), DE (Deutschland, Germany) and GB (Great Britain). However, unlike the variables depicting check-in method and payment method the number of options is large. The possibility remains to aggregate countries into continents, but Europe is overrepresented in the data. With over 180 different Points of Sale the decision was made not to represent each possibility but instead only keep columns for 8 most popular. All other countries are grouped into one variable called 'Other'. The doubt remains whether passenger behavior changes. For example, a Dutch passenger may book flights to Paris (France) from London (UK) while this had not been planned. The business travel model states nothing about Point of Sale as being a variable able to distinguish between passengers types.

Example of output with fictional data.

	Point of Sale			
Primary Key	**NL**	**DE**	**GB**	**Other**
Piggott James + DoB	0	0	7	1
Greeve Fai + DoB	4	0	0	1
Van Keulen Maurice + DoB	3	4	0	0
Amrit Chintan + DoB	2	0	0	0

A derivative variable counts the different points of sale per passenger.

Length of Stay

Redacted

Ticket lead-time

Redacted

Day of the week

However, PSQN does not contain the Date of Birth for passengers. As such a different key to combine PSQN and Altea had to be devised. The only viable key was based on last name, first letter of first name as no full first name was available, partial ticket number and PNR (Passenger Name Record). This key is also at times used within KLM when these databases are connected. However, a comparison of the population before and after based only on Altea variables show they are significantly different. As such a sample from the combined Altea_PSQN dataset cannot be used for clustering. Previous efforts by KLM to use this key must be called into question. In the last page of the appendix, this flawed result can be found.

Monet

The final database that is used is Monet, the revenue system for KLM.

Revenue

From Monet the average revenue for each subclass was extracted. These averages were used to calculate an estimate for the total revenue for each passenger over 2014. It is expected that as business travelers travel more often with high yield subclass of both Economy and Business class that passengers can be distinguished. A breakdown of the averages can be examined on the next page. Sub class G with an average value of only 36 euro per flight is an exception compared to other sub classes. It consists mostly of cheap.

Principal Components Analysis

Using SPSS to perform PCA analysis the following results were obtained. The Scree plot below shows for each component the corresponding Eigenvalue is an indicator to which extent the component explains variance in the data. Higher values are preferred. Components with Eigenvalues below or close to 1 have little variance and add little to any algorithm’s ability to infer a model or function. Such components should be removed from the dataset. Using the ‘elbow-method’ it can be determined that about 10 components have an Eigenvalue higher than 1. However, only the top 5 component have significant Eigenvalues

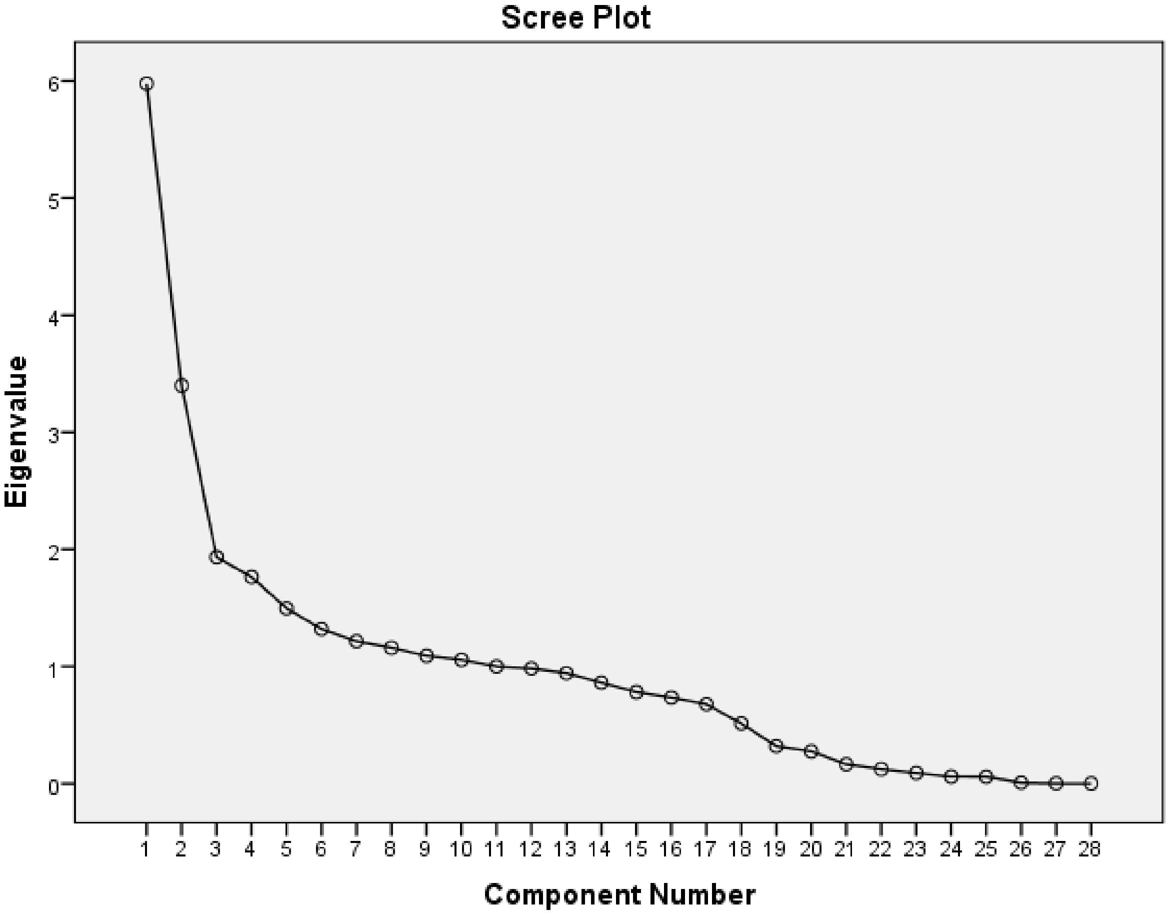


Figure 14 PCA results from SPSS

The 'Total Variance Explained' table shows that 11 components have an Eigenvalue higher than 1. Together they cumulatively explain 76.47 % of the variance found within the data set. To achieve a higher degree of variance explained more components have to be retained. However, such components only explain variance that is added by the component not the variables of the dataset. The top 17 components explain 92.1 % while the top 19 components explain the 95.9 %.

Total Variance Explained

Component	Initial Eigenvalues			Extraction Sums of Squared Loadings		
	Total	% of Variance	Cumulative %	Total	% of Variance	Cumulative %
1	5.978	21.349	21.349	5.978	21.349	21.349
2	3.399	12.138	33.487	3.399	12.138	33.487
3	1.933	6.905	40.392	1.933	6.905	40.392
4	1.764	6.299	46.691	1.764	6.299	46.691
5	1.496	5.343	52.034	1.496	5.343	52.034
6	1.320	4.713	56.747	1.320	4.713	56.747
7	1.215	4.339	61.086	1.215	4.339	61.086
8	1.159	4.138	65.225	1.159	4.138	65.225
9	1.091	3.898	69.122	1.091	3.898	69.122
10	1.056	3.771	72.893	1.056	3.771	72.893
11	1.000	3.573	76.466	1.000	3.573	76.466
12	.984	3.513	79.979			
13	.943	3.366	83.345			
14	.862	3.079	86.424			
15	.781	2.790	89.214			
16	.734	2.621	91.835			
17	.678	2.421	94.255			
18	.513	1.831	96.086			
19	.318	1.137	97.224			
20	.275	.981	98.205			
21	.165	.588	98.793			
22	.123	.440	99.233			
23	.090	.323	99.556			
24	.059	.211	99.767			
25	.058	.207	99.975			
26	.007	.024	99.998			
27	.000	.001	100.000			
28	8.543E-005	.000	100.000			

Extraction Method: Principal Component Analysis.

The components found with the PCA analysis do not directly correspond with the known variables of the data set. The 'Component Matrix' below shows the relationship between components and variables. Using this table a choice can be made which variables should be retained and which should be removed. For example, the variable frequency matches for more than 90 % with component 1 and should be kept as component 1 has the highest Eigenvalue (5.978). The variable DIAM does not closely match any components, no doubt due to the fact that few airline passengers are a member of this exclusive frequent flyer group.

Component Matrix Redacted

7. Results

In this chapter the results of the data clustering are explained in order to answer the sub research questions. The chapter will start by answering each of the sub questions before an attempt is made in the next chapter to answer the principal research question "Design a new airline market segment model with data clustering".

7.1. Unsupervised learning.

To answer the sub question "*Can clusters be associated with passenger segments and types?*" a number of Machine Learning algorithms are used to determine whether the dataset has underlying structures that can be used to answer business questions, discover market segments and specifically identify potential business travelers. The following algorithms were used: K-means, X-means,

K-Means algorithm

The first step of K-means clustering is to discover the optimal number of clusters. After this is discovered users of the algorithm are obligated to manually set this value each time it is used. As mentioned in the section of cluster validity (Chapter 4.5) the Sum of Squared Errors (SSE) is used. The following graph shows the value of SSE for every possible number of between 1 and 15 for a sample of the dataset. The SSE appreciable decreases when the number of clusters is increased to 2 and then 3. After the number of clusters is increased to 4 the SSE value ceases to decrease significantly. With K-means any number of clusters that is smaller than the size of the data points is possible, but such clustering won't describe anything appreciable about the underlying structure of the dataset. For K-means the optimal number of clusters is 4.

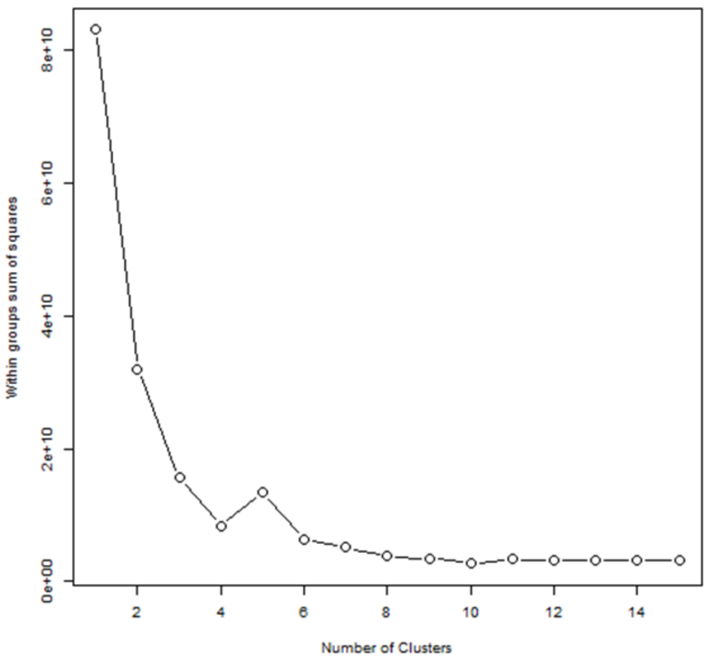


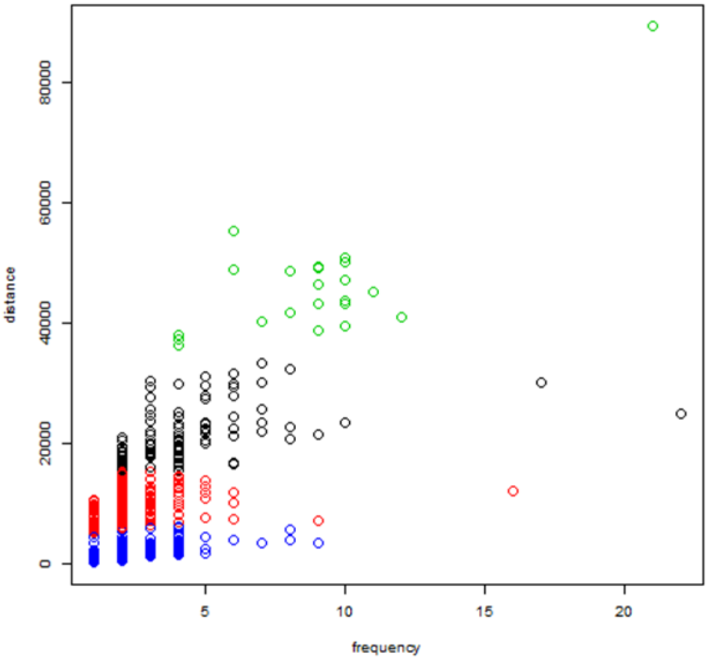
Figure 15 K-means: Sum of Square value per cluster count

The table below indicates that the clusters discovered have a much lower cohesion values compared to separation values. This is good, as it means that the clusters are very distinct from each other.

Test results	K-means			
Number of data points	999 (1 NA removed)			
Number of clusters	4			
Size of clusters	140 – 202 – 457 - 200			
Noise	0			
Cohesion	0	276	12921	7602
	276	0	5865	86
	12921	5865	0	272
	7602	86	272	0
Separation	0	14282	24224	18107
	14282	0	11971	5716
	24224	11971	0	7178
	18107	5716	7178	0
Average cohesion	2304			
Average Separation	12930			

Table 3 K-means cluster validation metrics

The plot below shows the result of the clustering algorithm across the space defined by frequency and distance. The data points are color coordinated using the results found by the K-means algorithm. This result confirms the suitability of the algorithm to define clusters interesting for market segmentation.



The next step is to evaluate the clustering results through the use of classification algorithms. Below the average F-measure for each classification algorithm is shown. With accuracy between 96.8 % and 99.9 % the results of K-means have proven to be very consistent. However, the PART algorithm, which is a variant of J48, has the highest value.

Algorithm	F-measure
J48	0.997
RandomTree	0.964
PART	0.999
BayesNet	0.968

Table 4 K-means classification results

A more detailed result for each classification algorithm can be found in the appendix.

Expectation–maximization algorithm

The EM-algorithm assigns a probability distribution to each instance in the dataset. Through cross-validation it finds the optimum number of clusters. To test this function the algorithm was also tested by specifying a priori that there should be only 2 clusters. The results of the clustering are shown below. As with K-means the algorithm discovers that the optimum number of clusters in the sample of the dataset is 4. The EM-algorithm prefers the log likelihood as it method of validation. However, for the use the algorithm remains a blackbox, validation of results can only be performed by considering the clustering results to be prior knowledge and use classification algorithms. The F-measure scores are below. The spread of the F-measure is narrower than with K-means, only Hierarchical clustering achieves a slightly narrower spread.

Algorithm	F-measure
J48	0.999
RandomTree	0.968
PART	0.998
BayesNet	0.973

Table 5 EM-algorithm F-measure score

X-means

The X-means algorithm has a wider spread for F-measure than K-means. With accuracy between 94.4 % and 99.8 % the results are very good and compare well in quality to K-means. However, the J48 algorithm has the highest value. The algorithm returns a result consisting of four clusters just like EM

Algorithm	F-measure
J48	0.998
RandomTree	0.951
PART	0.996
BayesNet	0.944

Table 6 X-means classification results

Hierarchical clustering results.

The hierarchical clustering algorithm uses the Ward metric to determine the optimal number of clusters. Unlike X-means the optimum number is 4, just like k-means. Below a graph, known as a dendrogram, show how the Hierarchical clustering algorithm has combined each entry of the dataset sample until the Ward criteria was met.

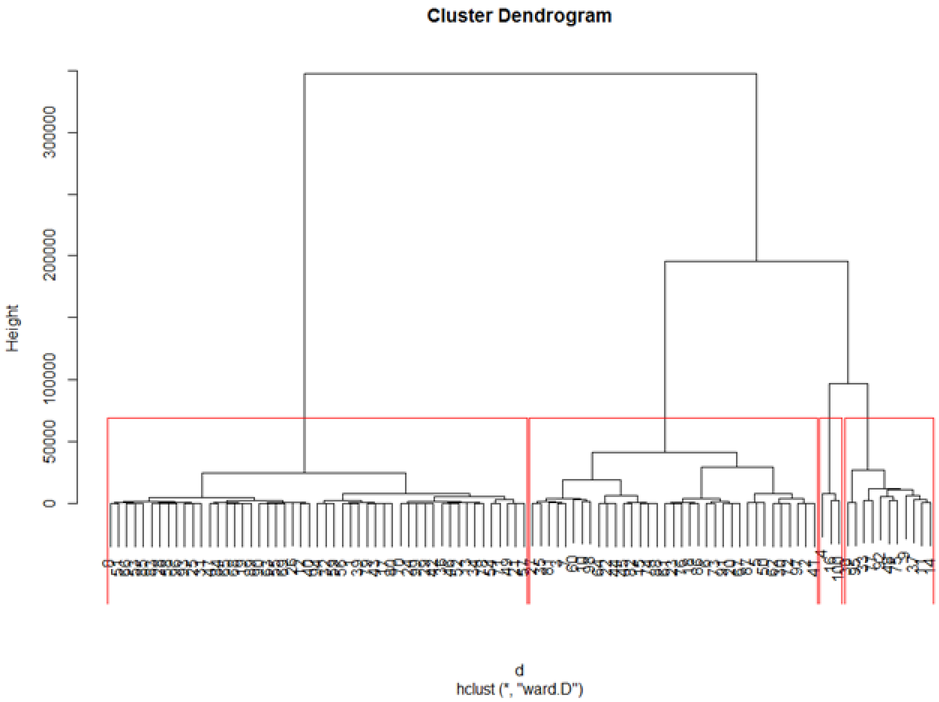


Figure 16 Dendrogram of Hierarchical Clustering Algorithm

Based on the F-measure score the spread in accuracy is narrower, with values between 96.9 % accuracy and 99.6 %. Again J48 and PART show the best result, both obtaining an accuracy of 96.9 %.

Algorithm	F-measure
J48	0.996
RandomTree	0.973
PART	0.996
BayesNet	0.969

Table 7 X-means classification results

Random clustering

To validate the correctness of using classification learning for evaluating unsupervised learning a random clustering result was also used. By assigning data points form the sample to randomly to one of four groups the expectation is that decision tree algorithm would achieve an average F-measure of 0.25 or 25 %.

Algorithm	F-measure
J48	0.252
RandomTree	0.254
PART	0.251
BayesNet	0.112

Table 8 Random clustering F-measure scores

The results of the random clustering conforms to expectations, though the results for BayesNet were unexpectedly low.

Conclusion

The classification algorithms have little difficulty accurately predicting the clusters values found with the unsupervised learning algorithms. K-means in particular, is known to be sensitive to outliers. Nonetheless its results are excellent. With a correct prediction rate of 99.9 % it is tied as the winner with EM-algorithm. This proves that validation of cluster results through classification is a viable alternative to the myriad of internal metrics of algorithms that are often hard to compare. Because of the narrower spread of F-measure found with EM-algorithm it is considered to provide the best clustering result. These were evaluated with expert stakeholders (see chapter 8) to answer the primary research question.

7.2. Semi-supervised learning.

To answer the sub question, “*Can an airline’s existing practice of customer segmentation be improved?”* all algorithms found in the WEKA package ‘collective-classification’ were used to determine whether it is possible to use semi-supervised learning to predict to which cluster new passenger records belong.

Algorithm	F-measure
Collective EM	0.75
Collective tree	0.702
Collective forest	0.743

Table 9 Semi-supervised F-measure score

The results from the algorithms proof that semi-supervised algorithms can predict new passenger records, but the level of accuracy is marginal. Collective EM-algorithm managed to cluster with a accuracy of 75 %.

Conclusion

The second research question is compared to the other two more abstract. However, it is a necessary step between clustering passengers into groups and creating behavior models. Take into account the Altea Departure Control system continues to grow with more systems added as well more passenger records added. Such growth can affect the future validity of cluster results as KLM market segment shift and passenger behavior changes. To perform the entire clustering process again would be costly and error prone. Essentially the clusters would have to labeled again and new decision trees created. With semi-supervised clustering a small amount of labeled passenger records is used to classify new passenger records. Each time this process is performed a new sample of labeled data is used. This process offers the possibility of shifting clusters without the need to re-label them.

7.3. Supervised learning.

The principle algorithm that is used to answer the third and final sub question is Decision tree learning. There are two models that are created and evaluated to successfully answer the research question: "*Can behavior of airline passengers be modeled?*"

1. Frequent Flyer model.
2. Corporate Flag model.

For each a static classification model is created and evaluated based on metrics such as F-score. In the next chapter, chapter 8, these models are evaluated for their utility with expert stakeholders

7.3.1. Frequent Flyer model.

Algorithm	**F-measure**
J48	0.998
RandomTree	0.984
PART	0.997
BayesNet	0.985

Table 10 frequent flyer F-measure scores

Conclusion

The J48 algorithm manages to achieve the highest average F-measure of all the algorithms tested, but the results for all are close. However, the Frequent Flyer tier levels that have relative fewer members are by all the algorithms harder to classify. This drop in accuracy especially noticeable with BayesNet. Taking equal samples from each tier level may be undesirable as there too few Platinum members. With Corporate Flag a similar result was seen and equal samples did make a difference (see next section).

Results Redacted

7.3.2. Corporate Flag model.

The first attempt of creating a decision tree model for Corporate flag failed because of the very low F-measure value found for those data points labeled as corporate versus non-corporate. The reason for the low score was because the number of data points are not equally distributed over both values. Non-Corporate Flag passengers outnumber the Corporate Flag passengers almost 10 to 1. The same tests were carried with values for both labels that are equal. The 465 Corporate Flag passenger records remained the same while from the 4300 Non-Corporate Flag passengers a random sample of 465 was extracted. The same classification tests were carried as with the first test series. The result of all 4 classification algorithms can be found below.

Algorithm	**F-measure**
J48	0.705
RandomTree	0.660

PART	0.702
BayesNet	0.719

Conclusion

The previous two test series have proven beyond doubt that the size of the subsets with regards to labels are an important factor in the accuracy of classification algorithms such as Decision Tree learning. Despite the fact that J48 obtained the second highest weighted average F-measure it is used to create a visual depiction (model) of Corporate Flag vs. Non-Corporate Flag passengers.

Results Redacted

Yes means passenger has corporate account

No means passenger does not have corporate account

8. Evaluation with stakeholders

In this chapter the results obtained with the clustering algorithm are interpreted in order to answer the primary research question, “Design a new airline market segment model with data clustering”. Most of the clustering algorithms agree that that the underlying data structure consists of 4 clusters. What follows is a description of each cluster that has been assembled through the stakeholder by confronting them with the result. The results of the Expectation-Maximization algorithm were used as reference.

The results were interpreted with the following expert stakeholders

Name	KLM position
Matthijs Neppelenbroek	Project Manager IMO sales
Arwed Wegscheid	Business Analyst
Maaike van der Horn	User insights manager

Results Redacted

9. Conclusion and limitations

Conclusion

The overall goal of this master thesis, to create a new market model and identify business travelers was attained with more success than anticipated. Clustering can yield fuzzy results and practically any dataset will cluster to some degree. Nonetheless, This result would not have been achieved without taking into care the myriad of problems that often plague Data analytics projects. Missing value, outliers and noise were a considerable hindrance. Ultimate connecting two databases, Altea Departure Control and PSQN booking system yielded a dataset that was skewered and whose cluster results could not be interpreted, but the effort proved it is a possible future avenue of research.

Through the results of this research effort KLM will be able to segment their customers and identify business travelers by applying the rules associated with the clusters directly into Altea. Thus passenger names can be extracted and potentially be targeted with better offers. Previous efforts by KLM to only using booking data with its flawed process of segregating passengers can thus be replaced. In anticipation of implementing a knowledge discovery system directly into Altea using Machine Learning algorithms this research effort has also proven that Semi-supervised clustering does work. This allows a small sample of previously clustered passenger records to be used to cluster new passenger into the previously established clusters. If such a system is implemented then no unsupervised clustering will have to be repeated, no expert knowledge will have to be consulted. Clusters will shift according to changing behavior of passengers and KLM's marketing efforts. Models of passenger behavior can then be extracted semi-regularly to extract relevant passenger groups.

Limitations

There are three threats to validity.

- The failure to properly merge the Altea and PSQN databases meant that variables such as length of stay at the destination and ticket lead time could not be included in the clustering efforts. These variable are arguably important to identify business passengers.
- The results of the clustering have been validated by corporate people stakeholders who may have their own interest to be either a sponsor or detractor of the findings of this research effort.
- The results of this study reflect the data from only one airline. Though the methodology can by be replicated for any other data set, or any other business with a large client pool.

References

Papers

1. Alexander, I., & Robertson, S. (2004). Understanding project sociology by modeling stakeholders. Software, IEEE, 21(1), 23-27
2. Alsabti, Khaled; Ranka, Sanjay; and Singh, Vineet, "An efficient k-means clustering algorithm" (1997). Electrical Engineering and Computer Science. Paper 43. http://surface.syr.edu/eecs/43
3. Alsultanny, Y. (2011). Selecting a suitable method of data mining for successful forecasting. Journal of Targeting, Measurement and Analysis for Marketing, 19(3), 207-225.
4. Arimond, G., & Elfessi, A. (2001). A clustering method for categorical data in tourism market segmentation research. Journal of Travel Research, 39(4), 391-397.
5. Arora, S., Raghavan, P., & Rao, S. (1998, May). Approximation schemes for Euclidean k-medians and related problems. In Proceedings of the thirtieth annual ACM symposium on Theory of computing (pp. 106-113). ACM.
6. Basu, S., Bilenko, M., & Mooney, R. J. (2004, August). A probabilistic framework for semi-supervised clustering. In Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining (pp. 59-68). ACM.
7. Becker, K., & Ghedini, C. (2005). A documentation infrastructure for the management of data mining projects. Information and Software Technology, 47(2), 95-111.
8. Behja, H., Marzak, A., & Trousse, B. (2012). Ontology-Based Knowledge Model for Multi-View KDD Process. lnternational Journal of Mobile Computing and Multimedia Communications (IJMCMC], 4(3), 21-33.
9. Bilenko, M., & Mooney, R. J. (2003, August). Adaptive duplicate detection using learnable string similarity measures. In Proceedings of the ninth ACM SIGKDD international conference on Knowledge discovery and data mining (pp. 39-48). ACM.
10. Brida, J. G., Scuderi, R., & Seijas, M. N. (2014). Segmenting Cruise Passengers Visiting Uruguay: a Factor–Cluster Analysis. International Journal of Tourism Research, 16(3), 209-222.
11. Caruana, R., & Niculescu-Mizil, A. (2006, June). An empirical comparison of supervised learning algorithms. In Proceedings of the 23rd international conference on Machine learning (pp. 161-168). ACM.
12. Chapman, P., Clinton, J., Kerber, R., Khabaza, T., Reinartz, T., Shearer, C., & Wirth, R. (2000). CRISP-DM 1.0 Step-by-step data mining guide.
13. Chaudhari, B., & Parikh, M. (2012). A Comparative Study of clustering algorithms Using weka tools. International Journal of Application or Innovation in Engineering & Management (IJAIEM), 1.
14. Cohn, D., Caruana, R., & McCallum, A. (2003). Semi-supervised clustering with user feedback. Constrained Clustering: Advances in Algorithms, Theory, and Applications, 4(1), 17-32.
15. Couvreur, C. (1997). The EM algorithm: A guided tour. In Computer Intensive Methods in Control and Signal Processing (pp. 209-222). Birkhäuser Boston.
16. Demiriz, A., Bennett, K. P., & Embrechts, M. J. (1999). Semi-supervised clustering using genetic algorithms. Artificial neural networks in engineering (ANNIE-99), 809-814.

17. Dempster, A. P., Laird, N. M., & Rubin, D. B. (1977). Maximum likelihood from incomplete data via the EM algorithm. Journal of the royal statistical society. Series B (methodological), 1-38.
18. Dolnicar, S. (2002). A review of data-driven market segmentation in tourism. Journal of Travel & Tourism Marketing, 12(1), 1-22.
19. EL MAHRSI, M. K., Etienne, C. O. M. E., Johanna, B. A. R. O., & Oukhellou, L. (2014, January). Understanding Passenger Patterns in Public Transit Through Smart Card and Socioeconomic Data: A case study in Rennes, France. In ACM SIGKDD Workshop on Urban Computing (p. 9p).
20. Ester, M., Kriegel, H. P., Sander, J., & Xu, X. (1996, August). A density-based algorithm for discovering clusters in large spatial databases with noise. In Kdd (Vol. 96, No. 34, pp. 226-231).
21. Fayyad, U., Haussler, D., & Stolorz, P. (1996). Mining scientific data. Communications of the ACM, 39(11), 51-57.
22. Fernández, D. M., & Wieringa, R. (2013). Improving requirements engineering by artefact orientation. In Product-Focused Software Process Improvement (pp. 108-122). Springer Berlin Heidelberg.
23. Grira, N., Crucianu, M., & Boujemaa, N. (2004). Unsupervised and semi-supervised clustering: a brief survey. A review of machine learning techniques for processing multimedia content, Report of the MUSCLE European Network of Excellence (FP6).
24. Hevner, A., March, S., Park, J., & Ram, S. (2004). Design science in information systems research. MIS quarterly, 28(1), 75-105.
25. Hourdakis, N., Argyriou, M., Petrakis, E. G., & Milios, E. E. (2010). Hierarchical Clustering in Medical Document Collections: the BIC-Means Method. JDIM, 8(2), 71-77.
26. Huls, C., Van Hillegersberg, J., Piggott, J.J.H. (2014). Micro-segmentation and personalization. Information Systems in the Financial Service Industry. Edited by Prof. Dr. Jos van Hillegersberg, Tim Bus & Hardwin Spenkelink. Chapter retrieved from ….
27. Jain, A. K. (2010). Data clustering: 50 years beyond K-means. Pattern Recognition Letters, 31(8), 651-666.
28. Jain, A. K., Murty, M. N., & Flynn, P. J. (1999). Data clustering: a review. ACM computing surveys (CSUR), 31(3), 264-323.
29. Jesri, S. H., Ahmadi, A., Karimi, B., & Shirazi, M. A. (2012). Hierarchical Data Clustering Model for Analyzing Passengers Trip in Highways. International Journal of Industrial Engineering, 23(4), 253-259.
30. Kanungo, T., Mount, D. M., Netanyahu, N. S., Piatko, C. D., Silverman, R., & Wu, A. Y. (2002). An efficient k-means clustering algorithm: Analysis and implementation. Pattern Analysis and Machine Intelligence, IEEE Transactions on, 24(7), 881-892.
31. Kishida, K. (2014). Empirical Comparison of External Evaluation Measures for Document Clustering by Using Synthetic Data (Vol. 2014, pp. 1-7). IPSJ SIG Technical Report.
32. Kitchenham, B. (2004). Procedures for performing systematic reviews. Keele, UK, Keele University, 33(2004), 1-26.
33. Klein, D., Kamvar, S. D., & Manning, C. D. (2002). From instance-level constraints to space-level constraints: Making the most of prior knowledge in data clustering.
34. Lloyd, S. (1982). Least squares quantization in PCM. Information Theory, IEEE Transactions on, 28(2), 129-137.

35. Maalouf, L., & Mansour, N. (2007, September). Mining airline data for CRM strategies. In Proceedings of the 7th WSEAS International Conference on Simulation, Modelling and Optimization, Beijing, China (pp. 345-350).
36. Meila, M., & Heckerman, D. (2013). An experimental comparison of several clustering and initialization methods. arXiv preprint arXiv:1301.7401.
37. Nadali, A., Kakhky, E. N., & Nosratabadi, H. E. (2011, April). Evaluating the success level of data mining projects based on CRISP-DM methodology by a Fuzzy expert system. In Electronics Computer Technology (ICECT), 2011 3rd International Conference on (Vol. 6, pp. 161-165). IEEE.
38. Pham, D. T., Dimov, S. S., & Nguyen, C. D. (2005). Selection of K in K-means clustering. Proceedings of the Institution of Mechanical Engineers, Part C: Journal of Mechanical Engineering Science, 219(1), 103-119.
39. Piggot, J., & Amrit, C. (2013). How Healthy Is My Project? Open Source Project Attributes as Indicators of Success. In Open Source Software: Quality Verification (pp. 30-44). Springer Berlin Heidelberg.
40. Raschka, S. (2014). Implementing a Principal Component Analysis (PCA) in Python step by step. Retrieved from http://sebastianraschka.com/Articles/2014_pca_step_by_step.html
41. Roy, S., & Bhattacharyya, D. K. (2005). An approach to find embedded clusters using density based techniques. In Distributed computing and internet technology (pp. 523-535). Springer Berlin Heidelberg.
42. Schumann, J.A. (2005) Data mining methodologies in educational organizations. PhD thesis, University of Connecticut, CT, USA.
43. Sharma, S., Osei-Bryson, K. M., & Kasper, G. M. (2012). Evaluation of an integrated Knowledge Discovery and Data Mining process model. Expert Systems with Applications, 39(13), 11335-11348.
44. Tukey, J. W. (1977). Exploratory data analysis.
45. Wagstaff, K., Cardie, C., Rogers, S., & Schrödl, S. (2001, June). Constrained k-means clustering with background knowledge. In ICML (Vol. 1, pp. 577-584).
46. Wieringa, R. (2010). Relevance and problem choice in design science. In Global Perspectives on Design Science Research (pp. 61-76). Springer Berlin Heidelberg.
47. Wieringa, R. (2012, January). Designing technical action research and generalizing from real-world cases. In Advanced Information Systems Engineering (pp. 697-698). Springer Berlin Heidelberg.
48. Wieringa, R., & Moralı, A. (2012). Technical action research as a validation method in information systems design science. In Design Science Research in Information Systems. Advances in Theory and Practice (pp. 220-238). Springer Berlin Heidelberg.
49. Xing, E. P., Jordan, M. I., Russell, S., & Ng, A. Y. (2002). Distance metric learning with application to clustering with side-information. In Advances in neural information processing systems (pp. 505-512).
50. Zhu, X. (2005). Semi-supervised learning literature survey.

Books

1. Bell, J. (2014). Machine Learning: Hands-On for Developers and Technical Professionals. John Wiley & Sons.
2. Marsland, S. (2014). Machine learning: an algorithmic perspective. CRC press.
3. Tan, P. N., Steinbach, M., & Kumar, V. (2006). Introduction to data mining (Vol. 1). Boston: Pearson Addison Wesley.
4. Wieringa, R. J. (2014). Design science methodology for information systems and software engineering. Springer.

Other source

1. CAPA, 2015. http://centreforaviation.com/analysis/air-france-klm-back-to-operating-loss-warns-lower-fuel-may-be-offset-by-low-unit-revenue--currency-210591. Retrieved on May 1st 2015.
2. SKIFT, 2015. http://skift.com/2015/01/18/air-france-klm-is-at-the-mercy-of-middle-east-and-low-cost-carriers/
3. Amadeus 2015, http://www.amadeus.com/airlineit/solutions/sol_1altea_5customer_1departure.html
4. Altea DC newsletter, July 20111.

Appendix

Table 1. Machine Learning algorithms, R packages and WEKA implementations used.

Algorithm model	**WEKA implementation**	**R package**
Unsupervised algorithms		
K-means		stats
K-medoid		cluster
Expectation Algorithm		EMCluster
Hierarchical clustering		pvclust
Density based clustering	DBSCAN, MakeDensityBasedClusterer, OPTICS	
X-means		cluster
Semi-supervised algorithms		
Expectation Algorithm	Collective EM	
Two	YATSI	
Decision Tree	Collective tree	
Random Forest	Collective forest	
Classifiers		
Decision Tree	RandomTree	
Naive Bayes classifier	BayesNet	
C4.5 (Decision Tree)	J48 and PART	

Table 2. List of variables found in the dataset or aggregated to the dataset.

Variable	**Data type**	**Description**
Frequency	Metric	Number of times a passenger has flown during 2014
Class	Nominal	Describes the class with which can be travelled (Business or Economy)
Subclass	Ordinal	Various book classes associated with Cabin class. 21 for within Europe and 19 for Intercontinental flights
Age	Ordinal	Age of passengers in years. Converted from Date of Birth.
Gender	Nominal	Gender of passenger. 0 for female, 1 for male and 2 for gender unknown
Distance	Metric	Distance traveled during 2014 in kilometers
Average distance	Metric	Average distance traveled per flight over 2014 . Distance
Day of the week	Nominal	The day of the week the flight took place.
Number of bags	Metric	Total number of bags a passenger has checked in during 2014
Average number of bags	Metric	Average number of bags check in per flight during 2014
Point of sale	Nominal	The country where the purchase of the ticket took place described as a two letter abbreviation. Examples; US for United States, NL for Netherlands etc.
Weight of bags	Metric	Total weight of bags checked in during 2014
Average weight of bags	Metric	Average weight of bags of a flight checked in during 2014
Tier level	Ordinal	Frequent flyer tier level obtained by passenger: Ivory, Silver, Gold, Platinum
Frequent flyer airline	Nominal	Passenger can be member of a frequent flyer program other than the one from KLM such as Delta or Air France.
Corporate program	Nominal	Boolean flag that show whether passenger has used a corporate contract
Length of stay	Metric	Average length of stay for a passenger at their destination.
Revenue	Metric	Total amount that a passenger has earned KLM during 2014. This does not include ancillaries.

Table 3. K-means result.

Below the detailed results are depicted. For each classification algorithm the average F-measure value is given while in the tables the F-measure value for each cluster is given.

J48 algorithm (implementation of C4.5)

Weighted Average F-measure is 0.997

One	Two	Three	Four	Classified as	F-measure
2706	0	2	0	**One**	0.999
0	476	0	1	**Two**	0.998
4	1	1504	0	**Three**	0.998
0	0	0	71	**Four**	0.993

RandomTree

Weighted Average F-measure is 0.964

One	Two	Three	Four	Classified as	F-measure
2686	1	21	0	**One**	0.992
0	440	26	11	**Two**	0.920
19	22	1466	2	**Three**	0.970
1	17	0	53	**Four**	0.774

PART

Weighted Average F-measure is 0.999

One	Two	Three	Four	Classified as	F-measure
2707	0	1	0	**One**	1.000
0	476	0	1	**Two**	0.997
1	2	1506	0	**Three**	0.998
0	0	0	71	**Four**	0.993

BayesNet

Weighted Average F-measure is 0.968

One	Two	Three	Four	Classified as	F-measure
2643	21	20	24	**One**	0.986
0	446	7	24	**Two**	0.922
9	19	1441	40	**Three**	0.968
0	4	0	67	**Four**	0.593

Table 4. X-means result

J48 algorithm (implementation of C4.5)

Weighted Average F-measure is 0.998

One	Two	Three	Four	Five	Classified as	F-measure
909	5	1	0	**0**	**One**	0.991
8	584	0	2	**0**	**Two**	0.986
0	0	476	0	**1**	**Three**	0.998
0	2	0	2706	**0**	**Four**	0.998
0	0	0	0	**71**	**Five**	0.993

RandomTree

Weighted Average F-measure is 0.951

One	Two	Three	Four	Five	Classified as	F-measure
842	47	21	4	**1**	**One**	0.896
48	519	5	22	**0**	**Two**	0.872
19	7	437	1	**13**	**Three**	0.920
13	23	4	2668	**0**	**Four**	0.988
1	1	6	0	**63**	**Five**	0.851

PART

Weighted Average F-measure is 0.996

One	Two	Three	Four	Five	Classified as	F-measure
909	5	1	0	**0**	**One**	0.990
8	584	0	1	**0**	**Two**	0.985
0	0	476	0	**1**	**Three**	0.998
0	3	0	2705	**0**	**Four**	0.999
0	0	0	0	**71**	**Five**	0.996

BayesNet

Weighted Average F-measure is 0.944

One	Two	Three	Four	Five	Classified as	F-measure
842	25	36	1	**11**	**One**	0.924
51	492	15	7	**29**	**Two**	0.870
13	0	441	0	**23**	**Three**	0.890
2	20	17	2643	**26**	**Four**	0.986
0	0	5	0	**66**	**Five**	0.584

Table 5. Hierarchical clustering.

J48 algorithm (implementation of C4.5)

Weighted Average F-measure is 0.996

One	Two	Three	Four	Classified as	F-measure
505	4	1	0	**One**	0.990
5	1456	0	2	**Two**	0.995
0	0	115	0	**Three**	0.996
0	5	3	2672	**Four**	0.999

RandomTree

Weighted Average F-measure is 0.973

One	Two	Three	Four	Classified as	F-measure
469	26	14	1	**One**	0.930
15	1419	3	26	**Two**	0.967
14	2	99	0	**Three**	0.853
1	26	1	2649	**Four**	0.990

PART

Weighted Average F-measure is 0.996

One	Two	Three	Four	Classified as	F-measure
505	4	1	0	**One**	0.990
5	1454	0	4	**Two**	0.994
0	0	115	0	**Three**	0.996
0	5	0	2672	**Four**	0.998

BayesNet

Weighted Average F-measure is 0.969

One	Two	Three	Four	Classified as	F-measure
484	3	23	0	**One**	0.931
32	1377	41	13	**Two**	0.966
3	0	112	0	**Three**	0.704
11	8	27	2631	**Four**	0.989

Table 6. classification results – Corporate flag.

The results below were discarded due to the very low F-measure score for data points labeled as corporate. Results in Table 8 shows the results for the adjusted data set.

Results Redacted

Table 7. Adjusted Corporate flag results.

Results Redacted

Table 8. Flying Blue membership results

Results Redacted

Flawed Altea_PSQN clustering Results Redacted